Introducción a la Estadística

Sheldon M. Ross

EDITORIAL REVERTÉ

INTRODUCCIÓN A LA ESTADÍSTICA

Sheldon M. Ross

1 Introducción a la Estadística

Estadística: arte de aprender a partir de los datos
Estadística descriptiva: describe y sintetiza los datos
Estadística inferencial: extrae conclusiones a partir de los datos
Población: conjunto de elementos de interés
Muestra: parte de la población a partir de la cual se obtienen los datos

2 Descripción de los conjuntos de datos

Tablas y gráficos de frecuencias y de frecuencias relativas
Histogramas
Diagramas de tallos y hojas
Gráficos de dispersión para datos apareados

3 Utilización de la Estadística para sintetizar conjuntos de dato

Media muestral: $\bar{x} = (\sum_{i=1}^{n} x_i)/n$

Mediana muestral: valor que ocupa la posición central

Varianza muestral: $s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)$

Desviación típica muestral: $s = \sqrt{s^2}$

Identidad algebraica: $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$

Regla empírica para los conjuntos de datos normales:
- aproximadamente el 68% de los datos cae dentro de $\bar{x} \pm s$
- aproximadamente el 95% de los datos cae dentro de $\bar{x} \pm 2s$
- aproximadamente el 99,7% de los datos cae dentro de $\bar{x} \pm 3s$

Coeficiente de correlación muestral:
$r = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})/[(n-1)s_x s_y]$

4 Probabilidad

$0 \le P(A) \le 1$
$P(S) = 1$, donde S es el conjunto de todos los valores posibles
$P(A \cup B) = P(A) + P(B)$, cuando A y B son disjuntos
Probabilidad del complementario: $P(A^c) = 1 - P(A)$
Regla de adición: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Probabilidad condicionada: $P(B|A) = P(A \cap B)/P(A)$
Regla de multiplicación: $P(A \cap B) = P(A)P(B|A)$
Sucesos independientes: $P(A \cap B) = P(A)P(B)$

5 Variables aleatorias discretas

Valor esperado (o media): $E[X] = \sum_{i=1}^{n} x_i P\{X = x_i\}$

$E[X + Y] = E[X] + E[Y]$
Varianza: $\mathrm{Var}(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2$
Desviación típica: $\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}$

$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ si X e Y son independientes
Variable aleatoria binomial:

$$P\{X = i\} = \frac{n!}{i!(n-i)!} p^i (1-p)^{n-i},\ i = 0, \ldots, n$$

$E[X] = np \quad \mathrm{Var}(X) = np(1-p)$

6 Variables aleatorias normales

Variable aleatoria normal X: caracterizada por $\mu = E[X]$, $\sigma = \mathrm{SD}(X)$
Variable aleatoria normal estándar Z: normal con $\mu = 0$, $\sigma = 1$
$P\{|Z| > x\} = 2P\{Z > x\},\ x > 0$
$P\{Z < -x\} = P\{Z > x\}$
z_α es tal que $P\{Z > z_\alpha\} = \alpha$
Si X es una normal, $Z = (X - \mu)/\sigma$ es una normal estándar
Propiedad aditiva: Si X e Y son normales independientes, $X + Y$ es también normal con media $\mu_x + \mu_y$, y varianza $\sigma_x^2 + \sigma_y^2$

7 Distribuciones de los estadísticos asociados al muestreo

$X_1, \ldots, X_n$ es una muestra procedente de una determinada población:
$E[X_i] = \mu$, $\mathrm{Var}(X_i) = \sigma^2$
$E[\bar{X}] = \mu$
$\mathrm{Var}(\bar{X}) = \sigma^2/n$
Teorema central del límite: $\sum_{i=1}^{n} X_i$ sigue aproximadamente, si n es grande, una normal con media $n\mu$ y desviación típica $\sigma\sqrt{n}$; equivalentemente a $\sqrt{n}(\bar{X} - \mu)/\sigma$ es aproximadamente una normal estándar.
Aproximación de la normal a la binomial: Si $np \ge 5$, $n(1-p) \ge 5$ se tiene que $[\mathrm{Bin}(n,p) - np]/\sqrt{np(1-p)}$ es aproximadamente una normal estándar.

8 Estimación

$\bar{X}$ es el estimador de la media poblacional μ.
$\hat{p}$, la proporción muestral de individuos que tienen una determinada propiedad, es un estimador de p, la proporción poblacional de individuos que tienen dicha propiedad.
S^2 estima σ^2 y S estima σ.
Estimador por intervalo a confianza $100(1-\alpha)\%$ para μ:
- Datos normales o n grande, σ conocido: $\bar{X} \pm z_{\alpha/2}\sigma/\sqrt{n}$
- Datos normales, σ desconocido: $\bar{X} \pm t_{n-1,\alpha/2} S/\sqrt{n}$

Intervalo de confianza a nivel $100(1-\alpha)\%$ para p: $\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$

9 Contraste de hipótesis estadísticas

H_0 = hipótesis nula: hipótesis que se desea contrastar.
Nivel de significación α: la (más alta posible) probabilidad de rechazar H_0 cuando es cierta
p valor: el menor nivel de significación al que H_0 sería rechazada.

Copyright © 2005 by Elsevier, Inc. All rights reserved.

Contrastes de hipótesis relativos a la media μ de una población

Supuesto: O bien la distribución es normal o bien el tamaño muestral n es grande

H_0	H_1	Estadístico del contraste TS	Contraste a nivel de significación α	p valor si TS = v
$\mu = \mu_0$	$\mu \neq \mu_0$	$\frac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma}$†	Rehazar H_0 si $\lvert TS\rvert \geq z_{\alpha/2}$	$2P\{z \geq \lvert v\rvert\}$
$\mu \leq \mu_0$	$\mu > \mu_0$	$\frac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma}$†	Rehazar H_0 si $TS \geq z_{\alpha}$	$P\{Z \geq v\}$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\frac{\sqrt{n}(\overline{X} - \mu_0)}{S}$	Rehazar H_0 si $\lvert TS\rvert \geq t_{n-1,\alpha/2}$	$2P\{T_{n-1} \geq \lvert v\rvert\}$
$\mu \leq \mu_0$	$\mu > \mu_0$	$\frac{\sqrt{n}(\overline{X} - \mu_0)}{S}$	Rehazar H_0 si $TS \geq t_{n-1,\alpha}$	$P\{T_{n-1} \geq v\}$

† *Supuesto*: σ conocido.
Observación: Para contrastar H_0: $\mu \geq \mu_0$, multiplique los datos por -1 y utilice lo anterior.

Contrastes de hipotesis relativos a p

(la proporción de elementos de una población que presentan cierta característica)

X es el número de elementos de una muestra de tamaño n que presentan la característica. B es una variable aleatoria binomial con parámetros n y p_0.

H_0	H_1	Estadístico del contraste TS	p valor si TS = x
$p \leq p_0$	$p > p_0$	X	$P\{B \geq x\}$
$p = p_0$	$p \neq p_0$	X	2 Min $\{P\{B \leq x\}, P\{B \geq x\}\}$

10 Contrastes de hipótesis relativas a dos poblaciones

Contrastes de hipótesis relativos a dos poblaciones cuando las muestras son independientes

La muestra X de tamaño n y la muestra Y de tamaño m son independientes.

H_0	H_1	Estadístico del contraste TS	Supuestos	Contraste a nivel de significación α	p valor si TS = v
$\mu_x = \mu_y$	$\mu_x \neq \mu_y$	$\frac{\overline{X} - \overline{Y}}{\sqrt{S_x^2/n + S_y^2/m}}$	n, m grandes	Rechazar H_0 si $\lvert TS\rvert \geq z_{\alpha/2}$	$2P\{Z \geq \lvert v\rvert\}$
$\mu_x \leq \mu_y$	$\mu_x > \mu_y$	$\frac{\overline{X} - \overline{Y}}{\sqrt{S_x^2/n + S_y^2/m}}$	n, m grandes	Rechazar H_0 si $TS \geq z_{\alpha}$	$P\{Z \geq v\}$
$\mu_x = \mu_y$	$\mu_x \neq \mu_y$	$\frac{\overline{X} - \overline{Y}}{\sqrt{S_p^2(1/n + 1/m)}}$	Poblaciones normales $\sigma_x = \sigma_y$	Rechazar H_0 si $TS \geq t_{n+m-2,\alpha/2}$	$2P\{T_{n+m-2} \geq \lvert v\rvert\}$
$\mu_x \leq \mu_y$	$\mu_x > \mu_y$	$\frac{\overline{X} - \overline{Y}}{\sqrt{S_p^2(1/n + 1/m)}}$	Poblaciones normales $\sigma_x = \sigma_y$	Rechazar H_0 si $TS \geq t_{n+m-2,\alpha}$	$P\{T_{n+m-2} \geq v\}$

$S_p^2 = \frac{n-1}{n+m-2} S_x^2 + \frac{m-1}{n+m-2} S_y^2 =$ estimador combinado de $\sigma_x^2 = \sigma_y^2$

Copyright © 2005 by Elsevier, Inc. All rights reserved.

Introducción a la **Estadística**

Sheldon M. Ross

EDITORIAL
REVERTÉ

Barcelona · Bogotá · Buenos Aires · Caracas · México

Título de la obra original:
Introductory Statistics. Second Edition

Edición original en lengua inglesa publicada por
Elsevier Inc. of 525B Street, Suite 1900, San Diego, CA 92101-4495, USA

Copyright © 2005, Elsevier Inc.

Edición en español:

© Editorial Reverté, S. A., 2007, 2014

Edición en papel:
ISBN: 978-84-291-5191-6

Edición e-book (PDF):
ISBN: 978-84-291-9424-1

Versión española traducida por:

Equipo de traducción coordinado por

Prof. Dr. Teófilo Valdés Sánchez
Departamento de Estadística e Investigación Operativa
Facultad de Matemáticas
Universidad Complutense de Madrid

Propiedad de:
EDITORIAL REVERTÉ, S. A.
Loreto, 13-15. Local B
08029 Barcelona. ESPAÑA
Tel: (34) 93 419 33 36
Fax: (34) 93 419 51 89
reverte@reverte.com
www.reverte.com

Reservados todos los derechos. La reproducción total o parcial de esta obra, por cualquier medio o procedimiento, comprendidos la reprografía y el tratamiento informático, y la distribución de ejemplares de ella mediante alquiler o préstamo públicos, queda rigurosamente prohibida sin la autorización escrita de los titulares del *copyright*, bajo las sanciones establecidas por las leyes.

1287

Sobre el autor

Sheldon M. Ross obtuvo su doctorado en Estadística por la Universidad de Stanford en 1968, tras ello se unió al Departamento de Ingeniería Industrial e Investigación Operativa de la Universidad de California, en Berkeley. Allí permaneció hasta el otoño de 2004, año en que obtuvo la plaza de Profesor de Ingeniería Industrial y de Sistemas en el Departamento Daniel J. Epstein de la Universidad del Sur de California. Ha publicado un gran número de artículos técnicos y de libros de texto en distintas áreas de Estadística y Probabilidad Aplicada. Entre sus libros de texto figuran *Un primer curso de Probabilidad* (sexta edición), *Introducción a los modelos de Probabilidad* (octava edición), *Simulación* (tercera edición) e *Introducción a la Probabilidad y a la Estadística para ingenieros y científicos* (tercera edición).

El profesor Ross es fundador de la revista *Probability in the Engineering and Informational Sciences*, revista de la que continúa siendo editor. Es miembro del Instituto de Estadística Matemática y ha recibido el Premio Humboldt para los Científicos Senior de Estados Unidos.

Para Rebecca y Elise

Prefacio

> Llegará un día en el que el razonamiento estadístico será tan necesario para el ciudadano como ahora lo es la habilidad de leer y escribir.
>
> H. G. Wells (1866-1946)

En el complicado mundo de hoy pocas cuestiones están claras y no sujetas a controversia. Para entender y formarnos una opinión sobre un tema debemos recoger información, es decir, datos. Y, para interpretar los datos, uno debe tener conocimientos de Estadística, que es el arte de sacar conclusiones a partir de los datos.

Este texto de estadística introductoria está dirigido a estudiantes universitarios de cualquier área. Se puede usar en un curso trimestral, semestral o anual. El único prerrequisito que exige es el de tener conocimientos de álgebra a nivel de enseñanza media. Su objetivo ha sido presentar los conceptos y las técnicas estadísticas de forma que pueda aportar a los estudiantes no solo cómo y cuándo se pueden utilizar los procedimientos estadísticos desarrollados, sino también para enseñarles a entender por qué se deben aplicar dichos procedimientos. Como resultado, se ha hecho un gran esfuerzo para explicar las ideas que sustentan los conceptos y las técnicas estadísticas presentadas. Los conceptos están razonados, ilustrados y explicados con la intención de potenciar la intuición del lector. Sólo cuando los estudiantes llegan a desarrollar un sentido o una intuición por la Estadística andan por el buen camino para extraer conclusiones a partir de los datos.

Para ilustrar las distintas aplicaciones de la Estadística y para mostrar a los estudiantes sus distintas perspectivas de uso, en este texto aparecen una amplia variedad de ejemplos y problemas de trabajo. La mayoría de ellos se refieren a cuestiones del mundo real, tales como el control de armas, los modelos de precios de acciones, algunas cuestiones de salud, los límites de edad para la conducción de vehículos, las edades de escolarización, los aspectos de política pública, el uso del casco, los deportes, la asignación de autor a obras literarias anónimas, el fraude científico, el beneficio potencial del consumo de vitamina C, entre otros. En muchos de ellos se utilizan datos que no sólo pertenecen al mundo real sino que, además, tienen interés por sí mismos. Los ejemplos se han planteado de forma clara y concisa, y se han incluido muchos problemas que incitan a pensar y que permiten desarrollar habilidades en la búsqueda de soluciones. Finalmente, algunos de los problemas han sido diseñados para que su solución quede abierta y para que sirvan de punto de arranque de posibles trabajos y proyectos por parte de los alumnos.

Algunas de las características especiales del texto

Introducción La primera sección de cada capítulo es una introducción, en la que se plantea una situación realista en términos estadísticos y con la que los estudiantes pueden tener una perspectiva sobre el contenido del capítulo.

La Estadística en perspectiva Éste es un apartado que aparece a lo largo de todo el texto para ilustrar con una aplicación del mundo real los conceptos y las técnicas estadísticos. Estas perspectivas están diseñadas para ayudar a los estudiantes a analizar e interpretar los datos y, también, a utilizar correctamente las distintas metodologías y técnicas estadísticas.

Datos reales A lo largo de todos los ejemplos, de las aplicaciones en el mundo real, de los problemas y los razonamientos incluidos en el texto, se utilizan conjuntos de datos reales para estimular el aprendizaje de los estudiantes. Estos conjuntos de datos suministran información para el estudio de temas de actualidad en una gran variedad de disciplinas, tales como la medicina y las ciencias de la salud, los deportes, los negocios y la educación.

Perspectivas históricas Estas secciones enriquecen el libro presentando los perfiles de estadísticos eminentes y de distintos hechos históricos; con ello se pretende que los estudiantes entiendan cómo ha evolucionado la Estadística a lo largo del tiempo.

Problemas/Problemas de repaso En este texto aparecen cientos de ejercicios al final de cada una de las secciones de los diferentes capítulos; igualmente, se han incluido problemas de repaso al final de cada capítulo. Muchos de estos problemas utilizan datos reales y están diseñados para valorar los conocimientos de los alumnos, tanto conceptuales como computacionales. Algunos problemas seleccionados se terminan de forma abierta y ofrecen una excelente oportunidad para organizar actividades y discusiones en grupo o para plantear trabajos individualizados a los alumnos.

Resumen/Términos clave En el resumen que hay al final de cada capítulo se presentan de forma concisa los conceptos y las fórmulas más importantes del mismo. Igualmente en cada capítulo se enumera un conjunto de términos clave y sus definiciones, a modo de glosario de trabajo.

Formulario Las fórmulas y las tablas más importantes que a menudo deben utilizar los estudiantes se incluyen en las contracubiertas inicial y final del libro. Pueden servir para consultas rápidas cuando se están realizando trabajos en casa o a la hora de preparar un examen.

Programas de ordenador En nuestro sitio web, www.reverte.com, se puede descargar el archivo STATCOMP.exe que contiene varios programas útiles para resolver problemas básicos de estadística. En el Apéndice E se da la lista completa de dichos programas. Para descargar el archivo, busque en nuestra página web el libro de Sheldom M. Ross y entre en Información adicional.

El libro de texto

El capítulo 1 sirve de presentación a la Estadística y se muestran sus dos ramas básicas. La primera, denominada Estadística Descriptiva, está relacionada con la recogida, la descrip-

ción y la síntesis de los datos. La segunda rama, denominada Estadística Inferencial, tiene por objeto la extracción de conclusiones a partir de los datos.

Los capítulos 2 y 3 están dedicados a la Estadística Descriptiva. En el capítulo 2 se analizan los métodos gráficos y tabulares que permiten presentar los conjuntos de datos. Se ve cómo una presentación efectiva de los conjuntos de datos facilita a menudo el descubrir algunas de sus características esenciales. El capítulo 3 muestra cómo se pueden resumir o sintetizar determinadas características de los conjuntos de datos.

Para poder sacar conclusiones a partir de los datos es preciso entender qué representan. Por ejemplo, habitualmente se asume que los datos constituyen una "muestra aleatoria procedente de una determinada población". Para entender exactamente lo que significa esta frase, y otras similares, es necesario tener ciertos conocimientos de Probabilidad; ésta es la finalidad del capítulo 4. El estudio de la Probabilidad es controvertido en una clase de introducción a la Estadística, porque suele ser bastante dificultosa para los estudiantes. Como resultado, ciertos libros de texto rebajan la importancia de este tema y lo presentan de forma superficial. Nosotros hemos elegido un enfoque diferente: hemos intentado concentrar la atención en los aspectos esenciales y presentarlos de forma clara y fácilmente entendible. Así pues, hemos tratado de forma breve, aunque cuidadosamente, los sucesos de un experimento, las propiedades de las probabilidades asignadas a dichos sucesos y los conceptos de probabilidad condicionada e independencia. El estudio de la Probabilidad continúa en el capítulo 5, donde se introducen las variables aleatorias discretas, y en el capítulo 6, dedicado a las variables aleatorias normales y a otras variables aleatorias continuas.

En el capítulo 7 se tratan las distribuciones de probabilidad de los estadísticos asociados al muestreo. También se analiza por qué la distribución normal tiene una importancia fundamental en la Estadística.

El capítulo 8 aborda el problema de utilizar los datos para estimar determinados parámetros de interés. Por ejemplo, podríamos desear estimar la proporción de personas que están en la actualidad a favor de una determinada medida política. Se estudian dos tipos de estimadores: el primero de ellos estima una magnitud de interés mediante un número (por ejemplo, se podría estimar que el 52% de la población está a favor de la medida política); el segundo tipo proporciona el estimador en la forma de un intervalo (por ejemplo, se podría estimar que el porcentaje de la población a favor de la medida política está comprendido entre el 49% y el 55%).

El capítulo 9 introduce un tema importante dedicado a los contrastes estadísticos de hipótesis, en el que se utilizan los datos para contrastar la plausibilidad de determinadas hipótesis. Por ejemplo, en un determinado contraste se podría rechazar la hipótesis de que más de un 60% de la población de votantes está a favor de una propuesta dada. También se incorpora el concepto de p-valor, que mide el grado de plausibilidad de una hipótesis una vez que se han observado los datos.

Mientras que los contrastes del capítulo 9 afectan a una sola población, los del capítulo 10 hacen referencia a dos poblaciones distintas. Por ejemplo, podemos estar interesados en contrastar si las proporciones de hombres y de mujeres a favor de una determinada propuesta coinciden.

Probablemente la técnica de inferencia estadística más extendida es la conocida como análisis de la varianza, que se expone en el capítulo 11. Dicha técnica nos permite contrastar hipótesis sobre parámetros que dependen de distintos factores. En este capítulo se analizan situaciones de análisis de la varianza unifactoriales y bifactoriales.

En el capítulo 12 se presenta la regresión lineal y cómo ésta se puede utilizar para relacionar el valor de una variable (digamos la altura de un hombre) con el de otra (la altura de su padre). Se analiza el concepto de regresión a la media y también se introduce y se explica detalladamente la falacia de la regresión. Se examina la relación entre regresión y correlación. Finalmente, en una sección opcional, se utiliza la regresión a la media junto con el teorema central del límite para presentar un razonamiento simple y original que explica por qué los conjuntos de datos biológicos suelen habitualmente seguir una distribución normal.

En el capítulo 13 se exponen los contrastes de bondad de ajuste, que sirven para contrastar si un determinado modelo propuesto es consistente con los datos. En este capítulo también se consideran poblaciones clasificadas según dos características y se muestra cómo contrastar si las características de un miembro de la población elegido aleatoriamente son independientes.

El capítulo 14 se centra en los contrastes de hipótesis no paramétricos, que son contrastes que se pueden usar en situaciones en las que los contrastes de los capítulos precedentes resultan inapropiados.

En el capítulo 15 se introduce el control de calidad, una técnica estadística clave relacionada con los procesos de transformación y producción.

Novedades de esta edición

Esta edición incluye muchos ejemplos y ejercicios nuevos y actualizados. Entre las secciones nuevas se hallan las siguientes:

- La sección 4.7, que es opcional y está dedicada a los principios de contaje.
- La sección 5.7, igualmente opcional, en la que se introducen las variables aleatorias de Poisson.
- La sección 12.10, en la que se evalúan los modelos de regresión mediante el análisis de los residuos.
- Las nuevas secciones del capítulo 15, sobre el control de calidad, en las que se introducen los gráficos de control de medias móviles ponderadas exponencialmente y de sumas acumuladas.

Contenido

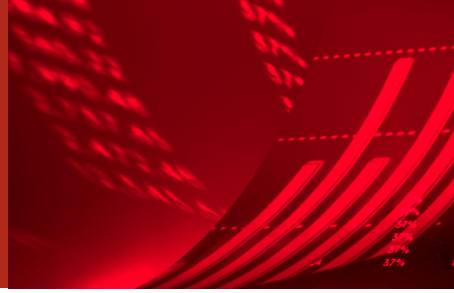
Agradecimientos

Nos gustaría dar las gracias a los siguientes revisores de la segunda edición:

James Wright, Universidad de Bucknell
Rodney Wong, Universidad de California en Berkeley
William Owen, Universidad de Case Western
Jaechoul Lee, Universidad de Boise State
Steven Garren, Universidad de James Madison
Pierre A. Grillet, Universidad de Tulane
Vincent Lariccia, Universidad de Delaware
John J. Deely, Universidad de Purdue
Cen-Tsong Lin, Universidad de Central Washington
Emily Silverman, Universidad de Michigan

Además queremos dar las gracias a Margaret Lin, Erol Pekoz, y a los siguientes revisores de la primera edición por sus útiles comentarios: William H. Beyer, Universidad de Akron; Patricia Buchanan, Universidad de Pennsylvania State; Michael Eurgubian, Santa Rosa Junior College; Larry Griffey, Florida Community College, Jacksonville; James E. Holstein, Universidad de Missouri; James Householder, Universidad de Humboldt State; Robert Lacher, Universidad de South Dakota State; Jacinta Mann, Seton Hill College; C. J. Park, Universidad de San Diego State; Ronald Pierce, Universidad de Eastern Kentucky; Lawrence Riddle, Agnes Scott College; Gaspard T. Rizzuto, Universidad de Southwestern Louisiana; Jim Robison-Cox, Universidad de Montana State; Walter Rosenkrantz, Universidad de Massachusetts, Amherst; Bruce Sisko, Belleville Area College; Glen Swindle, Universidad de California, Santa Barbara; Paul Vetrano, Santa Rose Junior College; Joseph J. Walker, Universidad de Georgia State; Deborah White, College of the Redwoods y Cathleen Zucco, LeMoyne College.

Sheldon M. Ross



Introducción a la **Estadística**

1

Introducción a la Estadística

Los estadísticos han invadido todas las ramas de la ciencia con una rapidez de conquista que sólo tiene como rivales a Atila, a Mahoma y al escarabajo de Colorado.

Maurice Kendall (estadístico británico)

Este capítulo introduce la materia objeto de la Estadística, el arte de aprender de los datos. Describe las dos ramas de la Estadística, la descriptiva y la inferencial. Se analiza la idea de aprender sobre una población a través de muestrear y estudiar a algunos de sus miembros. Finalmente se presentan algunos rasgos históricos.

1.1 Introducción

¿Es mejor que nuestros hijos sean escolarizados antes o después? Esta es una cuestión de interés para muchos padres y también para los gestores públicos. ¿Cómo se puede responder?

Inicialmente, parece razonable que nos planteemos esto a partir de nuestra propia experiencia y de algunas conversaciones mantenidas con los amigos. Sin embargo, si se quiere convencer a otras personas y obtener consensos, resulta necesario reunir algún tipo de información objetiva. Por ejemplo, en muchos Estados, los niños deben someterse a exámenes o a pruebas de conocimiento al final de su primer año de escolarización. Se pueden conseguir los resultados de los niños en dichas pruebas y analizarlos después para

ver si aparentemente existe una conexión entre la edad de escolarización y las calificaciones en las pruebas citadas. En realidad, tales estudios se han realizado y, por lo general, se ha concluido de ellos que los estudiantes de mayor edad han obtenido mejores calificaciones que los más jóvenes. Sin embargo, también se ha observado que los niños escolarizados a mayor edad son igualmente mayores a la hora de someterse a la prueba, y que este solo hecho por sí mismo podría ser la causa de sus más altas calificaciones. Por ejemplo, supongamos que los padres no enviaran a sus hijos a la escuela a los 6 años, sino un año más tarde. En este caso, puesto que durante ese año adicional los hijos aprenderían una gran cantidad de cosas en casa, tras su primer año de escuela, cuando realizaran la prueba podrían obtener calificaciones más altas que las que obtendrían si hubieran sido escolarizados un año antes, a los 6 años.

Un estudio reciente (tabla 1.1) ha intentado mejorar un trabajo anterior mediante el análisis del efecto que tenía la edad de escolarización sobre el número de años de escolarización. Los autores mantienen que este número de años mide mejor el éxito escolar que la calificación obtenida en el primer curso. A partir de los datos de los censos de 1960 hasta 1980, los autores concluyeron que la edad de escolarización incide muy poco sobre el número total de cursos completados. La tabla 1.1 incluye un compendio de los datos del estudio. La tabla muestra que, de los niños escolarizados en 1949, la mitad más joven (cuya edad media de escolarización fue de 6,29 años) se mantuvo escolarizada un promedio de 13,77 años, mientras que la otra mitad se mantuvo una media de 13,78 años.

Destacamos que no se ha pretendido presentar los anteriores datos como una prueba de que las edades de escolarización no afectan a su periodo de escolarización. Por el contrario, reflejan el enfoque moderno que tiene el uso de datos con respecto al análisis de situaciones complejas. En concreto, uno debe obtener información relevante, o datos, que han de ser descritos y analizados. Éste es el objetivo de la Estadística.

Tabla 1.1 Número total de años de escolarización con respecto a la edad de escolarización

	Mitad de los niños más jóvenes		Mitad de los niños de más edad	
Año de escolarización	**Edad media de escolarización**	**Número medio de años de escolarización**	**Edad media de escolarización**	**Número medio de años de escolarización**
1946	6,38	13,84	6,62	13,67
1947	6,34	13,80	6,59	13,86
1948	6,31	13,78	6,56	13,79
1949	6,29	13,77	6,54	13,78
1950	6,24	13,68	6,53	13,68
1951	6,18	13,63	6,45	13,65
1952	6,08	13,49	6,37	13,53

Fuente: ANGRIST J. y KRUEGER A., "The effect of age school entry on educational attainment: an application of instrumental variables with moments from two simples", en *Journal of the American Statistical Association*, 87, 18, 328-336.

1.2 La naturaleza de la Estadística

En el mundo de hoy, el que uno debe primero reunir datos para aprender sobre algo se ha convertido en un axioma. Por ejemplo, el primer paso para aprender sobre temas como

1. El estado actual de la economía.
2. El porcentaje de votantes a favor de una propuesta.
3. El número medio de kilómetros que puede recorrer un automóvil de nueva fabricación con un litro de gasolina.
4. La eficacia de un nuevo medicamento.
5. La utilidad de un nuevo método de enseñanza de lectura para niños de escuela elemental.

consiste en compilar los datos relevantes.

Definición

La *Estadística* es el arte de aprender a partir de los datos. Está relacionada con la recopilación de datos, su descripción subsiguiente y su análisis, lo que nos lleva a extraer conclusiones.

1.2.1 Obtención de datos

En ocasiones un análisis estadístico comienza con un conjunto de datos; por ejemplo, el gobierno habitualmente reúne datos sobre la tasa de desempleo y sobre el producto interior bruto. La Estadística se utiliza después para describir, clasificar y analizar esos datos.

En otras situaciones, los datos no están disponibles, y la Estadística se puede usar para diseñar un experimento apropiado para generar dichos datos. El experimento elegido dependería de la utilidad que se quiera obtener de los datos. Por ejemplo, si se acaba de desarrollar un medicamento reductor del colesterol y se quiere determinar su eficacia, se deben reclutar voluntarios y anotar sus niveles de colesterol. Después se les suministrará el medicamento durante cierto periodo de tiempo, y posteriormente se volverán a medir sus niveles de colesterol. Sin embargo, el experimento sería ineficaz si a *todos* los voluntarios reclutados se les suministrara el medicamento. Porque si fuera así, aunque los niveles de colesterol de todos los voluntarios se hubieran reducido significativamente, no estaría justificado concluir que las mejoras son debidas al medicamento en cuestión sino a alguna otra posibilidad. Es decir, está bien documentado el hecho de que cualquier medicación recibida por un paciente, tanto como si está o no directamente relacionada con la enfermedad sufrida, a menudo se traduce en mejoras en el estado del paciente. Esto se conoce como el *efecto placebo*, que no es tan sorprenderte como podría parecer inicialmente, puesto que la convicción que tiene el paciente de que se le está tratando de manera efectiva a menudo conduce a una reducción de su estrés, lo cual redunda en una mejora en su estado de salud. Adicionalmente, podrían haber existido otros factores, por lo general desconocidos, que influyeran sobre la reducción en los niveles de colesterol. Quizás el que la temperatura hubiera sido excepcionalmente cálida (o fría) podría haber hecho que los voluntarios estu-

vieran fuera de casa más o menos tiempo de lo habitual, lo que podría ser un factor determinante. Así pues, se ve que el experimento consistente en suministrar el medicamento a todos los voluntarios no está bien diseñado para generar datos a partir de los cuales se puedan sacar conclusiones acerca de la eficacia del medicamento.

Un experimento mejor intentaría neutralizar las posibles causas que afectan al nivel de colesterol, con excepción del medicamento. Una forma aceptada de conseguir esto consiste en dividir a los voluntarios en dos grupos: uno de ellos recibe el medicamento, mientras que el otro grupo recibe una pastilla (conocida como *placebo*) con la misma apariencia y sabor que el medicamento pero que no tiene ningún efecto fisiológico. Los voluntarios no deberían saber si se les está suministrando el medicamento o el placebo, y realmente sería mejor que tampoco lo supiera el personal médico que supervise el experimento, para que sus propias actitudes no jueguen papel alguno. Adicionalmente, es deseable que la división de voluntarios en dos grupos se haga de tal forma que ninguno de los grupos se vea favorecido en el sentido de que incluya a los "mejores" pacientes. Para conseguir esto, el procedimiento generalmente más aceptado consiste en que la división de voluntarios sea "aleatoria"; se entiende por este término que la división se haga de tal forma que todas las elecciones posibles de personas que compongan el grupo que recibe el medicamento sean igualmente probables. Al grupo que no recibe tratamiento alguno (los voluntarios que reciben el placebo) se le denomina grupo *de control*.

Una vez finalizado el experimento, se describirán los datos. Por ejemplo, se presentarían los niveles de colesterol de cada voluntario antes y después del experimento, y el experimentador anotaría para cada voluntario si éste ha recibido el medicamento o el placebo. Adicionalmente, se determinarían los valores sumariales, tales como la reducción media de colesterol de los miembros del grupo de control y de los miembros del grupo tratado con el medicamento.

Definición

La parte de la Estadística relacionada con la descripción y la clasificación de los datos se conoce con el nombre de *Estadística descriptiva*.

1.2.2 Estadística inferencial y modelos de probabilidad

Cuando se ha completado el experimento, y una vez que se han descrito y clasificado los datos, deseberíamos ser capaces de sacar conclusiones sobre la eficacia del medicamento. Por ejemplo, ¿se puede concluir que es efectivo como reductor de los niveles de colesterol en la sangre?

Definición

La parte de la Estadística relacionada con la extracción de conclusiones a partir de los datos se conoce con el nombre de *Estadística inferencial*.

Para poder sacar conclusiones a partir de los datos se ha de tener en cuenta el azar. Supongamos que la reducción media de colesterol es mayor para el grupo que recibió el medicamento que para el grupo de control. ¿Se puede concluir que ese resultado se debe al

medicamento, o es posible que éste sea realmente inefectivo y que la mejora se deba simplemente al azar? Por ejemplo, el hecho de que en 10 lanzamientos de una moneda resulten 7 caras no significa necesariamente que sea más probable la obtención de cara que la obtención de cruz en futuros lanzamientos. Realmente, podría tratarse de una moneda ordinaria y que, simplemente por azar, resultaran 7 caras en los 10 lanzamientos. (Sin embargo, si se hubiera obtenido 47 veces cara en 50 lanzamientos de la moneda, estaríamos bastante seguros de que no se trata de una moneda ordinaria.)

Para ser capaces de extraer conclusiones a partir de los datos suele ser necesario hacer determinadas hipótesis sobre las posibilidades (o *probabilidades*) de obtener los diferentes valores de los datos. La totalidad de esas hipótesis constituye el llamado *modelo de probabilidad* de los datos.

En ocasiones, la naturaleza de los datos sugiere cuál es la forma del modelo de probabilidad que se ha de elegir. Por ejemplo, supongamos que los datos consisten en las respuestas dadas por un grupo de individuos a una pregunta sobre si están a favor de una propuesta de reforma que afecta al bienestar social. Si el grupo fue seleccionado *aleatoriamente* parece razonable suponer que cada individuo consultado tenía una probabilidad p de decantarse a favor de la propuesta, donde p representa la proporción desconocida de ciudadanos en la población a favor de la propuesta. Se pueden utilizar los datos resultantes para hacer inferencias sobre p.

En otras situaciones, no resulta evidente cuál es el modelo de probabilidad adecuado para un determinado conjunto de datos. Sin embargo, una cuidadosa descripción y presentación de los datos nos permite inferir sobre un modelo razonable, que se puede intentar verificar posteriormente con el uso de datos adicionales.

Dado que la base de la inferencia estadística es la formulación de un modelo de probabilidad para describir los datos, para que ésta se pueda entender será necesario conocer previamente la teoría de la probabilidad. En otras palabras, la inferencia estadística comienza con la asunción de que ciertos aspectos importantes del fenómeno bajo estudio se pueden describir en términos de probabilidades, para luego llegar a hacer inferencias sobre estas probabilidades a través del uso de los datos.

1.3 Poblaciones y muestras

En Estadística, uno suele interesarse por obtener información sobre un conjunto total de elementos, al cual nos referiremos como la *población*. La población es a menudo demasiado grande para que se pueda examinar a cada uno de sus miembros. Por ejemplo, podría tratarse de todos los residentes de un determinado Estado, o de todos los aparatos de televisión producidos por una determinada compañía en el último año, o del conjunto de hogares de una comunidad dada. En tales casos se intenta aprender sobre la población eligiendo a un subgrupo de sus elementos, que luego será examinado. Este subgrupo de la población se llama *muestra*.

Definición

El conjunto total de elementos en los que estamos interesados se llama *población*.

Un subgrupo de la población que será estudiado en detalle se llama *muestra*.

Para que la muestra proporcione información sobre la población total, deberá ser, en algún sentido, representativa de dicha población. Por ejemplo, supongamos que estamos interesados en aprender sobre la distribución de edades de los residentes de una ciudad y que obtenemos las edades de las 100 primeras personas que entran en una determinada biblioteca de la ciudad. Si la edad media de esas 100 personas es de 46,2 años, ¿podemos concluir justificadamente que este valor coincide aproximadamente con la edad media de toda la población? Posiblemente no, porque seguro que se podría argüir que la muestra elegida no es en este caso representativa de la población total, ya que generalmente son los estudiantes jóvenes y los ciudadanos mayores quienes frecuentan la citada biblioteca, en mayor medida que las personas que están en edad laboral. Se ha de tener en cuenta que el término *muestra representativa* no significa que la distribución de los individuos de la muestra coincida exactamente con la de la población total, sino que la muestra ha sido elegida de forma que todos los elementos de la población tengan la misma probabilidad de pertenecer a la muestra.

En ciertas situaciones, como en el caso de la biblioteca, se nos suministra una muestra y debemos decidir si es una muestra razonablemente representativa de la población total. En la práctica, una muestra dada no puede, por lo general, considerarse representativa de una población, a menos que la muestra haya sido elegida de forma aleatoria. Esto ocurre porque cualquier procedimiento no aleatorio para seleccionar una muestra suele proporcionar resultados sesgados a favor de algunos valores de datos y en contra de otros.

Definición

Una muestra de k miembros de una población se dice que es una *muestra aleatoria*, en ocasiones llamada *muestra aleatoria simple*, si los miembros son elegidos de tal forma que todas las posibles elecciones de los k miembros son igualmente probables.

Así, aunque pueda parecer paradójico, es más factible obtener una muestra representativa si sus miembros son elegidos de forma totalmente aleatoria, sin considerar *a priori* qué elementos deben ser elegidos. En otras palabras, no se ha de intentar deliberadamente elegir la muestra de forma que nos parezca que contiene, por ejemplo, la misma proporción por sexo o por profesión que la población total. Por el contrario, todo ello se ha de dejar al "azar" para, a partir de la muestra, obtener aproximaciones de las proporciones correctas en la población. La mecánica habitual de selección de muestras aleatorias implica el uso de números aleatorios, que aparecen en el Apéndice C.

Una vez elegida la muestra aleatoria, se puede utilizar la inferencia estadística para sacar conclusiones sobre la población total mediante el estudio de los elementos de la muestra.

*1.3.1 Muestreo aleatorio estratificado

Un método más sofisticado que el muestreo aleatorio simple es el *muestreo aleatorio estratificado*. Este tipo de muestreo requiere mayor información sobre la población que el mues-

* El asterisco señala temas opcionales que no se tratarán en los apartados siguientes.

treo aleatorio simple. Supongamos que un instituto de secundaria tiene: 300 estudiantes en el primer curso, 500 en el segundo y 600 en los cursos tercero y cuarto. Supongamos que, para conocer la respuesta de los estudiantes a una propuesta militar que afecta a los jóvenes de 18 años, se decide entrevistar en detalle a 100 estudiantes. En lugar de elegir aleatoriamente a 100 estudiantes entre los 2000 existentes, en el muestreo estratificado se calcula cuántos estudiantes se van a elegir de cada curso. Puesto que la proporción de estudiantes en el primer curso es de $300/2000 = 0{,}15$, en una muestra estratificada el porcentaje será el mismo; en consecuencia, se seleccionarán en la muestra $100 \times 0{,}15 = 15$ estudiantes de primer curso. De igual forma, se seleccionarán $100 \times 0{,}25 = 25$ estudiantes de segundo curso, $100 \times 0{,}30 = 30$ de tercer curso y a 30 de cuarto curso. Cada uno de estos conjuntos de estudiantes se seleccionará aleatoriamente entre los alumnos de cada curso.

En otras palabras, en este tipo de muestreo, primero la población se *estratifica* en subpoblaciones, y luego los elementos se eligen aleatoriamente dentro de cada subpoblación. Como resultado, la proporción de elementos muestrales que pertenecen a cada subpoblación coinciden con las proporciones de la población total. La estratificación es particularmente efectiva para averiguar las proporciones medias en la población con respecto a una pregunta de interés, cuando existen diferencias significativas entre las subpoblaciones. Por ejemplo, en la encuesta anterior, los estudiantes del último curso, siendo mayores, podrían verse más inmediatamente afectados por la propuesta militar que los estudiantes de los cursos anteriores. Por consiguiente, en las consecuencias de la propuesta podrían existir diferencias entre los cursos, y la estratificación sería efectiva para conocer la receptibilidad media a la misma.

1.4 Breve historia de la Estadística

La recopilación sistemática de datos económicos y de población se inició en Venecia y Florencia, las ciudades-Estado italianas, durante el Renacimiento. El término *Estadística*, derivado de la palabra *Estado*, se utilizó entonces para referirse a la obtención de datos de interés estatal. Esta idea de recopilación de datos se extendió desde Italia a otros países de la Europa occidental. De hecho, durante la primera mitad del siglo XVI, era habitual que los gobiernos europeos obligaran a las parroquias a que registraran los nacimientos, los matrimonios y las defunciones. Debido a las muy escasas condiciones de salud pública, las estadísticas referidas a estos hechos tenían un especial interés.

Las altas tasas de mortalidad en Europa antes del siglo XIX se debieron primordialmente a epidemias, guerras y hambruna. Entre las epidemias, las peores eran las plagas. Desde la Peste negra de 1348, frecuentemente, se sucedieron plagas durante cerca de 400 años. En 1562, como forma de conseguir que la corte real se trasladara al campo, la ciudad de Londres comenzó a publicar los datos de mortalidad. Inicialmente, esos datos listaban los lugares de defunción y si las muertes habían sido causadas por dicha plaga. Desde 1625, esta información se extendió a todas las causas de defunción.

En 1662 el comerciante inglés John Graunt publicó un libro titulado *Observaciones naturales y políticas hechas a partir de los datos de mortalidad*. La tabla 1.2, que incluye el número total de fallecimientos en Inglaterra y el número de ellos causados por la peste en cinco años diferentes de epidemia, está sacada del citado libro.

Tabla 1.2 Total de fallecimientos en Inglaterra

Año	Entierros	Muertes por peste
1592	25 886	11 503
1593	17 844	10 662
1603	37 294	30 561
1625	51 758	35 417
1636	23 359	10 400

Graunt utilizó los datos de mortalidad de Londres para estimar la población de la ciudad. Por ejemplo, para estimar la población de Londres en 1660, Graunt muestreó los hogares de ciertas parroquias (o suburbios) de Londres y descubrió que, en media, se producían 3 defunciones por cada 88 habitantes. Dividiendo entre 3, observó que en media había una muerte por cada 88/3 habitantes. Puesto que las cifras de mortalidad de Londres recogían 13 200 muertes en Londres en el año en cuestión, Graunt estimó que la población de Londres era de aproximadamente

$$13\,200 \cdot \frac{88}{3} = 387\,200$$

Graunt utilizó este estimador para pronosticar cuál era el número total de habitantes de Inglaterra. Apuntó en su libro que tales cifras serían de interés para los gobernantes del país, como indicadores tanto del número de hombres que podrían movilizarse en el ejército como de los que que podrían contribuir con los impuestos.

Graunt también usó las cifras de mortalidad de Londres –así como algunas inteligentes intuiciones sobre qué enfermedades mataban a qué personas y a qué edades– para inferir las tasas de defunción por edad. (Recuerde que los datos de mortalidad listaban solamente las causas y los lugares de defunción, no las edades de los fallecidos.) Graunt utilizó esa información para confeccionar tablas que representaban las proporciones poblacionales de muerte por distintas clases de edad. La tabla 1.3 es una de las tablas de mortalidad de Graunt. Por ejemplo, incluye que, de cada 100 nacimientos, 36 personas morirían antes de alcanzar la edad de 6 años, 24 morirían con una edad comprendida entre 6 y 15 años, y así sucesivamente.

Los estimadores de Graunt sobre la mortalidad por edades fueron de gran interés en los negocios de gestión de pensiones. Éstos se diferenciaban de los seguros de vida en que la gente aportaba una cantidad establecida como inversión y, a su cuenta, recibía una cantidad regular de por vida.

Los trabajos de Graunt sobre tablas de mortalidad inspiraron la aportación de Edmund Halley en 1693. Halley, descubridor del cometa que lleva su nombre (y también el hombre que más apoyó, tanto psicológica como económicamente, la publicación del famoso libro *Principia Mathematica* de Isaac Newton), utilizó las tablas de mortalidad para calcular las probabilidades que una persona de cualquier edad tenía de sobrepasar otra edad distinta. Halley ejerció gran influencia para convencer a las aseguradoras del momento de que los seguros tenían que depender de las edades de los asegurados.

Tras Graunt y Halley, la recopilación de datos se incrementó de manera continuada durante todo el siglo XVII y hasta bien entrado el XVIII. Por ejemplo, la ciudad de París

Tabla 1.3 Tabla de mortalidad de Graunt

Edad de muerte	Muertes por cada 100 nacimientos
0–6	36
6–16	24
16–26	15
26–36	9
36–46	6
46–56	4
56–66	3
66–76	2
≥76	1

Nota: Las clases se acercan al valor de la derecha, pero no lo incluyen. Por ejemplo, 0-6 incluye las edades de 0 a 5 años.

empezó a registrar cifras de mortalidad en 1667; y en 1730 registrar las edades de muerte era una práctica común en toda Europa.

El término *Estadística*, que se utilizó hasta el siglo XVIII como una abreviatura de la ciencia descriptiva de los Estados, se identificó cada vez más, en el siglo XIX, con las cifras cuantitativas. Hacia 1830, en Francia e Inglaterra, el término ya fue usado de forma general como sinónimo de la *ciencia numérica* de la sociedad. Este cambio de significado se debió a que, desde 1800, los gobiernos de Europa occidental y de Estados Unidos comenzaron a recopilar y publicar sistemáticamente una gran cantidad de registros de censos y de otros tipos de tablas.

Aunque a lo largo del siglo XIX la teoría de la probabilidad había sido desarrollada por matemáticos tales como Jacob Bernoulli, Karl Friedich Gauss y Pierre Simon Laplace, su aplicación al estudio de hechos estadísticos fue casi inexistente, ya que la mayor parte de los estadísticos sociales de la época se contentaban con dejar que los datos hablaran por sí mismos. En particular, en esa época los estadísticos no estaban interesados en sacar inferencias a partir de individuos, más bien se centraban en la sociedad en su totalidad. Por consiguiente, no estaban preocupados por el muestreo sino que intentaban obtener censos de la población al completo. Como resultado, la inferencia probabilística sobre la población a partir de muestras era prácticamente desconocida en las estadísticas sociales del siglo XIX.

No fue hasta finales de este siglo cuando los estadísticos empezaron a preocuparse por inferir conclusiones a partir de los datos numéricos. El movimiento comenzó con los trabajos de Francis Galton sobre el análisis de la influencia de la herencia a través de la utilización de técnicas que actualmente se conocen como análisis de regresión y correlación (véase el capítulo 12), que alcanzaron su mayor auge con los trabajos de Karl Pearson. Éste, que desarrolló los contrastes de bondad de ajuste (véase el capítulo 13), fue el primer director del laboratorio Galton, fundado por Galton en 1904. Allí, Pearson lideró un programa de investigación con el objetivo de desarrollar nuevos métodos en los que la Estadística se utilizaba con fines inferenciales. Su laboratorio potenció que investigadores provenientes de distintas áreas de la ciencia y la industria aprendieran los métodos estadísticos que podían tener aplicación en sus campos. Uno de los primeros estudiantes que

acudió a su laboratorio fue W. S. Gosset, químico de formación, que mostró su devoción por Pearson y publicó sus propios trabajos bajo el seudónimo de *Student*. Existe una famosa leyenda que mantiene que Gosset temía publicar bajo su propio nombre por miedo a que, en la fábrica de cerveza Guinnes, sus jefes descubrieran que uno de sus químicos realizaba investigaciones estadísticas. Gosset es famoso por su desarrollo de los contrastes de la t (véase el capítulo 9).

Dos de las áreas más importantes de aplicación de la estadística a principios del siglo XX fueron la biología y la agricultura; todo ello gracias al interés de Pearson y de otros investigadores de su laboratorio, y también gracias a los logros del científico inglés Ronald A. Fisher. La teoría de la inferencia desarrollada por estos investigadores pioneros –y otros, entre los que se encuentran Egon Pearson, hijo de Karl Pearson, y el estadístico matemático polaco Jerzy Neyman– fue lo suficientemente general como para tratar una amplia gama de problemas cuantitativos prácticos. Como resultado, tras los primeros años del siglo XX, aumentó muy rápidamente el número de personas, procedentes de la ciencia, los negocios y la administración, que empezaron a considerar la estadística como una herramienta capaz de suministrar soluciones cuantitativas a una gran variedad de problemas científicos y prácticos.

En la actualidad, podemos encontrar las ideas estadísticas en muchos ámbitos. La Estadística descriptiva puede verse en periódicos y revistas. La *Inferencia Estadística* se ha hecho indispensable en las investigaciones médicas y de salud pública, en la investigación de mercados y en el control de calidad, en la educación, la contabilidad, la economía, en la predicción meterológica o de las votaciones, y en muestreos, deportes, seguros, en juegos de apuestas y en cualquier tipo de investigación que pretenda ser científica. Hoy en día, la Estadística se ha convertido realmente en una pieza integrante de nuestra herencia intelectual.

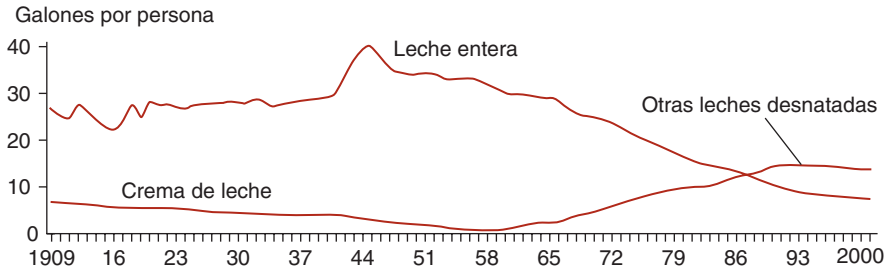
Problemas

1. Este problema tiene relación con la tabla 1.1.

   (a) ¿En qué año hubo la mayor diferencia entre el número medio de cursos finalizados por los estudiantes que comenzaron su escolarización antes o después?

   (b) ¿Existieron más años en los que el promedio de cursos completados por el grupo de estudiantes que comenzaron más jóvenes fue mayor que el del grupo de estudiantes que comenzaron tarde o sucedió al contrario?

2. El siguiente gráfico muestra los consumos de leche en Estados Unidos desde 1909 hasta 2000. ¿Qué conclusión general se puede extraer?



3. Los siguientes datos muestran los porcentajes de fumadores adultos en Estados Unidos, clasificados por sexo y nivel educacional, entre los años 1999 y 2002.

   (a) ¿En qué grupos ha existido una reducción sistemática?

   (b) ¿Podría decirse que existe una tendencia general?

Consumo de cigarrillos en Estados Unidos (% de adultos fumadores)

	1999	2000	2001	2002
Total	**25,8**	**24,9**	**24,9**	**26,0**
Sexo				
Hombre	28,3	26,9	27,1	28,7
Mujer	23,4	23,1	23,0	23,4
Educación				
No graduado en secundaria	39,9	32,4	33,8	35,2
Graduado en secundaria	36,4	31,1	32,1	32,3
Con algunos cursos universitarios	32,5	27,7	26,7	29,0
Graduado unversitario	18,2	13,9	13,8	14,5

4. Intentando determinar la eficacia de un medicamento nuevo, un investigador médico ha comenzado con el contraste del medicamento frente a un placebo. Para asegurarse de que los dos grupos de pacientes voluntarios –aquellos que reciben el medicamento y los que reciben el placebo– son lo más parecidos posible, el investigador ha decidido no basarse en el azar sino que, por el contrario, ha analizado detalladamente a los voluntarios y luego él mismo ha elegido los grupos. ¿Es aconsejable este procedimiento? ¿Por qué sí? o ¿por qué no?

5. Explique por qué es importante que un investigador que intenta estudiar la utilidad de un nuevo medicamento no conozca qué pacientes son tratados con el medicamento y cuáles están recibiendo el placebo.

6. Se va a celebrar una votación la semana próxima y se pretende predecir, mediante la selección de una muestra de votantes, si ganará el candidato republicano o el candidato demócrata. ¿Cuál de los siguientes métodos de selección permite obtener una muestra significativa?

   (a) Seleccionar a toda la gente en edad de votar que asiste a un partido de baloncesto universitario.

   (b) Seleccionar a todas las personas en edad de votar que salen de un restaurante de moda de la ciudad.

   (c) Obtener una copia del censo de votantes, elegir 100 nombres aleatoriamente y entrevistarlos.

   (d) Utilizar los resultados de un programa de televisión en el que se pide a los espectadores que llamen por teléfono y comuniquen su elección.

   (e) Elegir nombres de la guía telefónica y llamarles posteriormente.

7. El procedimiento indicado en el problema 6(e) condujo a una predicción desastrosa en las elecciones presidenciales de 1936, en las que Franklin Roosevelt venció a Alfred Landon por una mayoría aplastante. El *Literary Digest* había vaticinado la victoria de Landon. Esta revista había basado su predicción en las preferencias manifestadas por una muestra de votantes obtenida a partir de las listas de propietarios de vehículos y de teléfonos.

   (a) ¿Por qué cree que la predicción del *Literary Digest* resultó tan errónea?

   (b) ¿Ha cambiado algo, desde 1936 hasta la actualidad, que nos permita creer que el procedimiento empleado por *Literary Digest* funcionaría hoy mejor?

8. Un investigador pretende descubrir la edad media de defunción actual de los habitantes de Estados Unidos. Para obtener datos, lee las columnas de fallecimientos publicadas por el periódico *New York Times* durante 30 días y anota las edades de los fallecidos. ¿Piensa que este procedimiento permite obtener una muestra representativa?

9. Si en el problema 8 la edad media de los fallecidos registrados fue de 82,4 años, ¿qué conclusión se podría sacar?

10. Para determinar el porcentaje de fumadores de una ciudad se ha decidido muestrear a diferentes personas en uno de los siguientes lugares de la ciudad:

    (a) el interior de una piscina (b) una pista de una bolera

    (c) un centro comercial (d) una biblioteca

    ¿En cuál de estos lugares potenciales es más probable que se obtenga una aproximación razonable a la proporción buscada? ¿Por qué?

11. Una universidad decide llevar a cabo una encuesta entre sus graduados recientes para obtener información sobre sus salarios anuales. Se seleccionó aleatoriamente a 200 graduados recientes y se les enviaron cuestionarios referidos a sus empleos actuales. Sin embargo, de esos 200, sólo 86 rellenaron y devolvieron el cuestionario. Supongamos que la media de los salarios reportados fue de 75 000 dólares.

    (a) ¿Sería correcto que la universidad pensara que 75 000 dólares es una buena aproximación al nivel medio de salarios de todos sus graduados? Explique el razonamiento en que se basa la respuesta.

    (b) Si su respuesta en (a) es no, ¿puede pensar en un conjunto de condiciones referidas al grupo de cuestionarios devueltos para el cual los 75 000 dólares sería una buena aproximación?

12. En un artículo se reportó que una encuesta sobre la ropa que llevaban por los peatones muertos por atropellos nocturnos había revelado que el 80% de las víctimas llevaba ropas oscuras, mientras que el restante 20% vestía ropas claras. La conclusión a la que se llegaba en el artículo fue que era más seguro llevar ropas claras por la noche.

    (a) ¿Está justificada esta conclusión? Explíquelo.

    (b) Si su respuesta en (a) fue no, ¿qué otra información se necesitaría antes de sacar una conclusión final?

13. Critique el método de Graunt para estimar la población de Londres. ¿Qué hipótesis implícita se está asumiendo?

14. Las cifras de mortalidad de Londres registraban 12 246 muertes en 1658. Suponiendo que una encuesta sobre las parroquias de Londres mostró que, grosso modo, un 2% de la población había fallecido en dicho año, utilice el método de Graunt para estimar la población de Londres en 1658.
15. Suponga que usted es un vendedor de planes de pensiones en 1662, año en el que se publicó el libro de Graunt. Explique cómo habría usado los datos sobre las edades en las que se producían los fallecimientos.
16. Si se basa en la tabla 1.2, ¿cuál de los cinco años de peste parece haber sido el más severo? Explique su razonamiento.
17. Basándose en la tabla de mortalidad de Graunt:
    (a) ¿Qué proporción de bebés sobrevivió a la edad de 6 años?
    (b) ¿Qué proporción de bebés sobrevivió a la edad de 46 años?
    (c) ¿Qué proporción murió entre las edades de 6 y 36 años?
18. ¿Por qué piensa que el estudio de la Estadística es importante en sus áreas de interés? ¿Cómo cree que puede utilizarla en su trabajo futuro?

La cambiante definición de la Estadística

La Estadística tiene el objetivo de realizar una representación fiable de un Estado en una época determinada. (Quetelet, 1849)

La Estadística es la única herramienta mediante la cual se puede conseguir una apertura en la formidable espesura de dificultades que entorpece el camino de aquellos que estudian la Ciencia del hombre. (Galton, 1889)

La Estadística puede considerarse (i) como el estudio de las poblaciones, (ii) como el estudio de las variaciones y (iii) como el estudio de los métodos de reducción de datos. (Fisher, 1925)

La Estadística es la disciplina científica relativa a la recopilación, el análisis y la interpretación de datos obtenidos mediante la observación o la experimentación. Tiene una estructura coherente basada en la *Teoría de la Probabilidad* e incluye muchos procedimientos diferentes que contribuyen a la investigación y el desarrollo en todas las ramas de la Ciencia y la Tecnología. (E. Pearson, 1936)

La Estadística es el nombre de la ciencia que trata de llevar a cabo inferencias bajo situaciones de incertidumbre; para ello, usa los números para averiguar cuestiones relativas a la naturaleza y la experiencia. (Weaver, 1952)

La Estadística se caracteriza en el siglo XX como una herramienta matemática para analizar datos experimentales u observados. (Ross, 2005)

Términos clave

Estadística: Arte de aprender de los datos.

Estadística descriptiva: Parte de la Estadística que trata con la descripción y la clasificación de los datos.

Estadística inferencial: Parte de la Estadística relativa a la extracción de conclusiones a partir de los datos.

Modelo de probabilidad: Hipótesis matemáticas relativas a verosimilitud de los distintos valores de los datos.

Población: Conjunto de elementos de interés.

Muestra: Subgrupo de la población que va a ser estudiado.

Muestra aleatoria de tamaño *k*: Muestra seleccionada de tal forma que todos los subgrupos de tamaño *k* tienen la misma probabilidad de ser seleccionados.

Muestra aleatoria estratificada: Muestra obtenida tras dividir la población en distintas subpoblaciones y elegir después muestras aleatorias en cada una de subpoblaciones citadas.

2

Descripción de los conjuntos de datos

Los números constituyen el único lenguaje universal.

Nathaniel West

La gente que no cuenta no cuenta.

Anatole France

En este capítulo se aprenderán métodos para presentar y describir conjuntos de datos. Se introducirán distintos tipos de tablas y gráficos, que permitirán ver fácilmente las características clave de un conjunto de datos.

2.1 Introducción

Es muy importante que los resultados numéricos de cualquier estudio se presenten en forma clara y concisa, de modo que rápidamente se pueda tener una idea de las características esenciales de los datos. Esto es particularmente necesario cuando se trata de un amplio conjunto de datos, como frecuentemente ocurre en las encuestas o en los experimentos controlados. Realmente, una presentación efectiva de los datos a menudo revela con rapidez elementos tales como su categoría, su grado de simetría, lo concentrados o dispersos que están, dónde se concentran, etcétera. En este capítulo se tratarán distintas técnicas de presentación de datos, tanto tabulares como gráficas.

Las tablas y los gráficos de frecuencias que se presentan en la sección 2.2 incluyen una gran variedad de tablas y gráficos –gráficos de línea, gráficos de barras, y gráficos de polígono– que son útiles para describir conjuntos de datos que tienen un relativamente pequeño número de valores distintos. A medida que el número de valores distintos crece, éstos van dejando de ser efectivos, y es más conveniente dividir los datos en clases distintas para considerar solamente el número de valores que pertenecen a cada una de las clases. Esto se hace en la sección 2.3, donde se estudian los histogramas, un tipo de gráfico de barras que resulta de representar gráficamente las frecuencias de las clases. En la sección 2.4 se estudia una variación del histograma, conocida como gráfico de tallos y hojas, variación que utiliza los propios valores de los datos para representar los tamaños de las clases. En la sección 2.5 se considera la situación en la que los datos corresponden a pares de valores, como por ejemplo la población y la tasa de criminalidad de distintas ciudades, y se introduce el diagrama de dispersión como método efectivo de presentación de dichos datos. Finalmente, en la sección 2.6 se exponen algunos comentarios históricos.

2.2 Tablas y gráficos de frecuencias

Los siguientes datos representan los días de baja por enfermedad en las últimas 6 semanas de un grupo de 50 trabajadores de una cierta compañía.

2, 2, 0, 0, 5, 8, 3, 4, 1, 0, 0, 7, 1, 7, 1, 5, 4, 0, 4, 0, 1, 8, 9, 7, 0,
1, 7, 2, 5, 5, 4, 3, 3, 0, 0, 2, 5, 1, 3, 0, 1, 0, 2, 4, 5, 0, 5, 7, 5, 1

Puesto que este conjunto de datos contiene un número relativamente pequeño de valores diferentes, conviene representarlo en una *tabla de frecuencias*, la cual incluye cada valor distinto junto con su frecuencia de ocurrencia. La tabla 2.1 es la tabla de frecuencias de los datos anteriores. En dicha tabla, la columna de frecuencias representa el número de ocurrencias de cada valor distinto del conjunto de datos. Observe que la suma de todas las frecuencias es 50, el número total de datos observados.

Ejemplo 2.1 Utilice la tabla 2.1 para contestar a las preguntas siguientes:

(a) ¿Cuántos trabajadores han estado de baja por enfermedad al menos 1 día por enfermedad?

(b) ¿Cuántos trabajadores han estado de baja entre 3 y 5 días, ambos inclusive?

(c) ¿Cuántos trabajadores han han estado de baja más de 5 días?

Tabla 2.1 Tabla de frecuencias de los días de baja por enfermedad

Valor	Frecuencia	Valor	Frecuencia
0	12	5	8
1	8	6	0
2	5	7	5
3	4	8	2
4	5	9	1

Solución

(a) Puesto que 12 de los 50 trabajadores no estuvieron ningún día de baja, la respuesta es $50 - 12 = 38$.

(b) La respuesta es la suma de las frecuencias de los valores 3, 4 y 5; esto es, $4 + 5 + 8 = 17$.

(c) La respuesta es la suma de las frecuencias de los valores 6, 7, 8 y 9. Por tanto, la respuesta es $0 + 5 + 2 + 1 = 8$. ■

2.2.1 Gráficos de líneas, gráficos de barras y polígonos de frecuencias

Se pueden mostrar gráficamente los datos de una tabla de frecuencias mediante un *gráfico de líneas*, en el que los valores sucesivos se representan sobre el eje horizontal y sus correspondientes frecuencias se representan mediantea la altura de una línea vertical. La figura 2.1 muestra el gráfico de líneas para los datos de la tabla 2.1.

En ocasiones, las frecuencias se representan no se representan mediante líneas sino mediante barras de una cierta anchura. Estos gráficos, llamados *gráficos de barras*, se utilizan muy a menudo. La figura 2.2 presenta un gráfico de barras que se corresponde con los datos de la tabla 2.1.

Otro tipo de gráfico utilizado para representar una tabla de frecuencias es el *polígono de frecuencias*, en el que se muestran gráficamente las frecuencias de los diferentes valores de los datos y luego se conectan los puntos del gráfico mediante líneas rectas. La figura 2.3 presenta el polígono de frecuencias de los datos de la tabla 2.1.

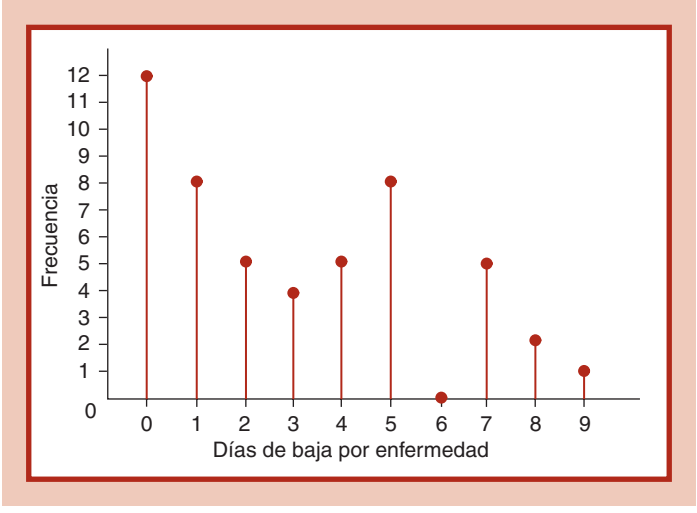


Figura 2.1 Gráfico de líneas.

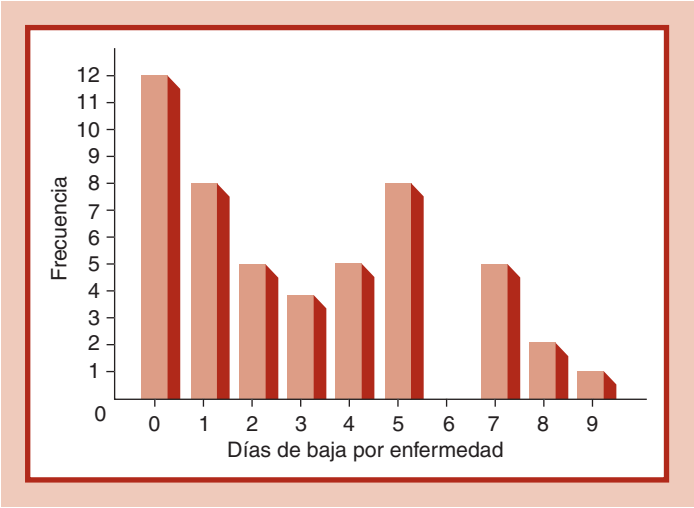


Figura 2.2 Gráfico de barras.

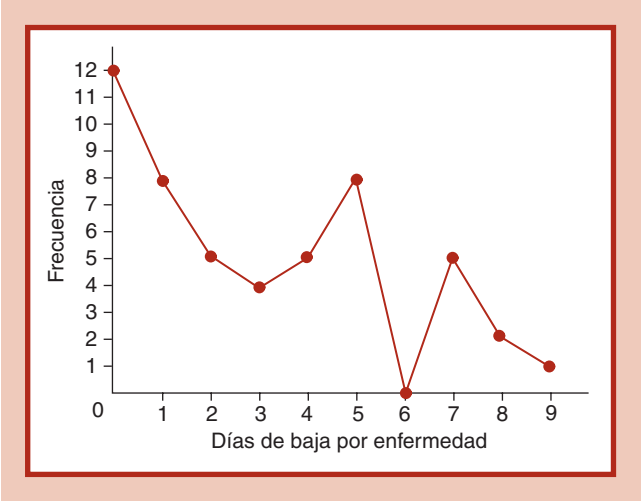


Figura 2.3 Polígono de frecuencias.

Se dice que un conjunto de datos es simétrico con respecto al valor x_0 si las frecuencias de los valores $x_0 - c$ y $x_0 + c$ son iguales para todo c. Es decir, para cada constante c, existe el mismo número de datos con un valor igual a c unidades por debajo de x_0 que con un valor igual a c unidades por encima de x_0. El conjunto de datos reflejado en la tabla de frecuencias de la tabla 2.2 es simétrico con respecto al valor $x_0 = 3$.

Los datos "próximos" a ser simétricos se dice que son *aproximadamente simétricos*. La forma más fácil de determinar si un conjunto de datos es aproximadamente simétrico consiste en representarlos gráficamente. La figura 2.4 incluye tres gráficos de barras: un con-

Tabla 2.2 Tabla de frecuencias de un conjunto de datos simétrico

Valor	Frecuencia	Valor	Frecuencia
0	1	4	2
2	2	6	1
3	3		

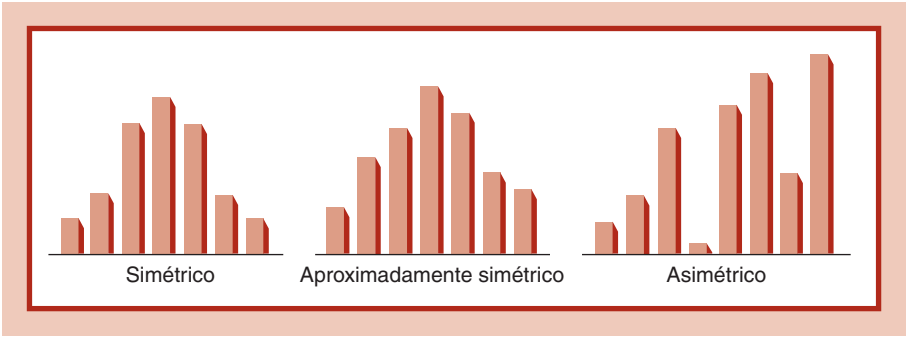


Figura 2.4 Gráfico de barras y simetría.

junto de datos simétrico, un conjunto de datos aproximadamente simétrico, y el último, un conjunto de datos asimétrico.

2.2.2 Gráficos de frecuencias relativas

En ocasiones, es más conveniente considerar y representar gráficamente las frecuencias *relativas* que las frecuencias *absolutas* de los datos. Si f representa la frecuencia de ocurrencia del valor x, se puede mostrar gráficamente la *frecuencia relativa* f/n frente a x, donde n representa el número total de observaciones del conjunto de datos. Para los datos de la tabla 2.1, $n = 50$ y las frecuencias relativas vienen reflejadas en la tabla 2.3. Observe que, mientras que la suma de la columna de frecuencias es igual al número total de observaciones del conjunto de datos, la suma de la columna de frecuencias relativas es 1.

En la figura 2.5 se presenta un polígono de frecuencias para las citadas frecuencias relativas. Un gráfico de frecuencias relativas tiene la misma apariencia que el gráfico análogo de frecuencias absolutas, aunque los valores del eje vertical se han dividido entre el número total de observaciones del conjunto de datos.

Para construir una tabla de frecuencias relativas de un conjunto de datos

Ordene el conjunto de datos de forma creciente en valores. Determine los valores distintos y sus frecuencias de ocurrencia. Liste los citados valores distintos junto con sus frecuencias f y sus frecuencias relativas f/n, donde n es el número total de observaciones del conjunto de datos.

Tabla 2.3 Frecuencias relativas de los datos de días de baja por enfermedad, $n = 50$.

Valor x	Frecuencia f	Frecuencia relativa f/n
0	12	$\frac{12}{50} = 0{,}24$
1	8	$\frac{8}{50} = 0{,}16$
2	5	$\frac{5}{50} = 0{,}10$
3	4	$\frac{4}{50} = 0{,}08$
4	5	$\frac{5}{50} = 0{,}10$
5	8	$\frac{8}{50} = 0{,}16$
6	0	$\frac{0}{50} = 0{,}00$
7	5	$\frac{5}{50} = 0{,}10$
8	2	$\frac{2}{50} = 0{,}04$
9	1	$\frac{1}{50} = 0{,}02$

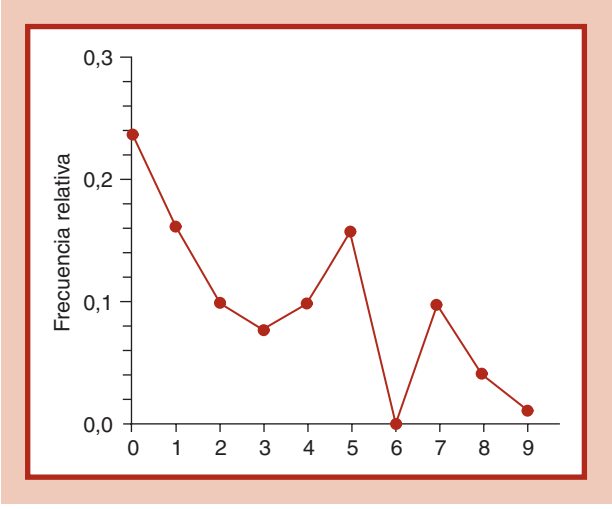


Figura 2.5 Polígono de frecuencias relativas.

Ejemplo 2.2 El Torneo de Maestros de Golf se juega cada año en Augusta (Georgia), en el Club Nacional de Golf. Para analizar las puntuaciones que han tenido los vencedores de este torneo, se han registrado las puntuaciones ganadoras desde 1968 hasta 2004.

Vencedores del Torneo de Maestros de Golf

Año	Vencedor	Puntuación	Año	Vencedor	Puntuación
1968	Bob Goalby	277	1987	Larry Mize	285
1969	George Archer	281	1988	Sandy Lyle	281
1970	Billy Casper	279	1989	Nick Faldo	283
1971	Charles Coody	279	1990	Nick Faldo	278
1972	Jack Nicklaus	286	1991	Ian Woosnam	277
1973	Tommy Aaron	283	1992	Fred Couples	275
1974	Gary Player	278	1993	Bernhard Langer	277
1975	Jack Nicklaus	276	1994	J.M. Olazabal	279
1976	Ray Floyd	271	1995	Ben Crenshaw	274
1977	Tom Watson	276	1996	Nick Faldo	276
1978	Gary Player	277	1997	Tiger Woods	270
1979	Fuzzy Zoeller	280	1998	Mark O'Meara	279
1980	Severiano Ballesteros	275	1999	J.M. Olazabal	280
1981	Tom Watson	280	2000	Vijay Singh	278
1982	Craig Stadler	284	2001	Tiger Woods	272
1983	Severiano Ballesteros	280	2002	Tiger Woods	276
1984	Ben Crenshaw	277	2003	Mike Weir	281
1985	Bernhard Langer	282	2004	Phil Nickelson	279
1986	Jack Nicklaus	279			

(a) Organice el conjunto de puntuaciones ganadoras mediante una tabla de frecuencias relativas.

(b) Represente estos datos mediante un gráfico de barras de frecuencias relativas.

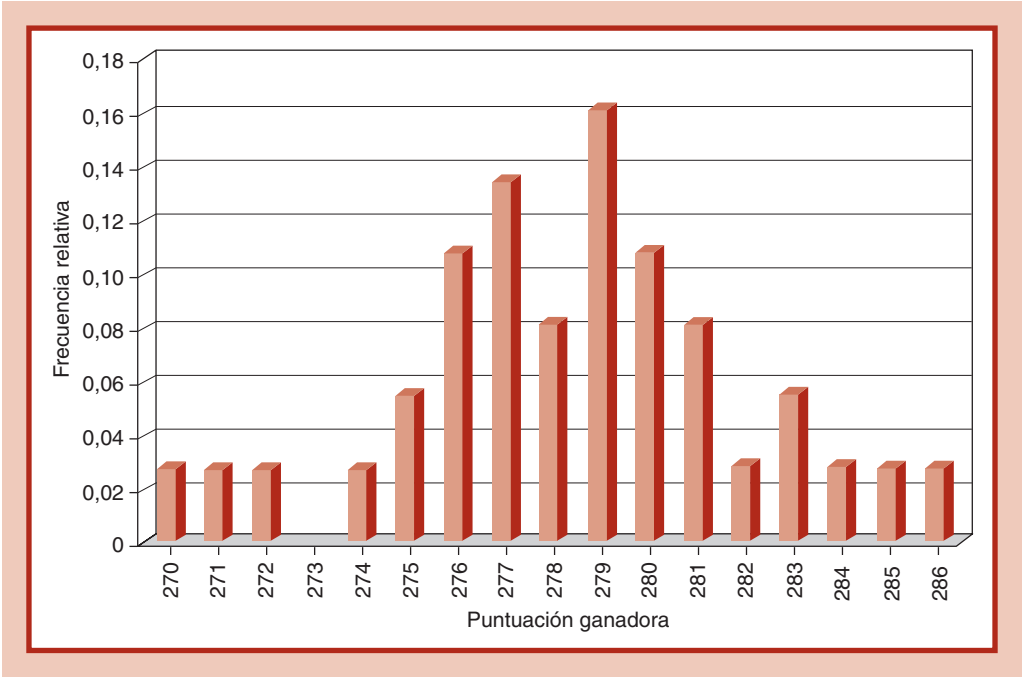
Solución

(a) Las 37 puntuaciones ganadoras varían desde la más baja de 270 hasta la más alta de 289. Su tabla de frecuencias relativas es la siguiente:

Puntuación ganadora	Frecuencia f	Frecuencia relativa $f/37$
270	1	0,027
271	1	0,027
272	1	0,027
274	1	0,.027

Puntuación ganadora	Frecuencia *f*	Frecuencia relativa *f*/37
275	2	0,054
276	4	0,108
277	5	0,135
278	3	0,081
279	6	0,162
280	4	0,108
281	3	0,081
282	1	0,027
283	2	0,054
284	1	0,027
285	1	0,027
286	1	0,027

(b) Un gráfico de barras de los datos anteriores es el siguiente:



2.2.3 Gráficos de tarta

Los *gráficos de tarta* se suelen utilizar para representar frecuencias relativas cuando los datos no son numéricos. Se construye un círculo que luego se divide en sectores, uno por cada valor diferente de datos. El área de cada sector, con la que se pretende representar la

Tabla 2.4 Armas utilizadas en los asesinatos.

Tipo de arma	Porcentaje de asesinatos causados con esta arma
Pistola de mano	52
Cuchillo	18
Escopeta	7
Rifle	4
Herramienta personal	6
Otras	13

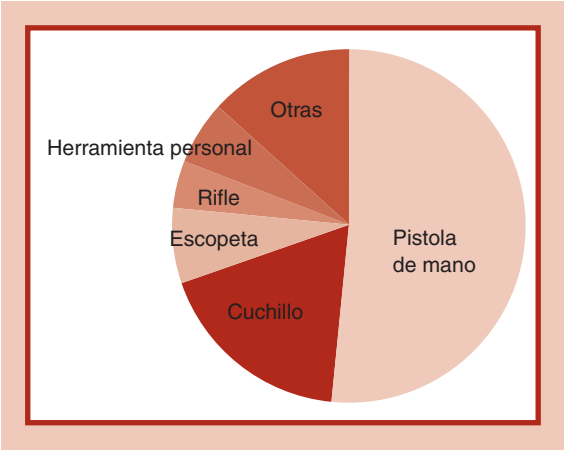


Figura 2.6 Gráfico de tarta.

frecuencia relativa de un valor, se determina como sigue. Si la frecuencia relativa del valor es f/n, el área de su sector debe coincidir con la fracción f/n del área total del círculo. Por ejemplo, los datos de la tabla 2.4 muestran las frecuencias relativas a las armas usadas en los asesinatos producidos en una gran ciudad durante 1985. Estos gráficos se representan mediante un gráfico de tarta en la figura 2.6.

Si un determinado valor tiene una frecuencia relativa f/n, su sector correspondiente puede obtenerse con la selección de un ángulo igual a $360f/n$ grados. Por ejemplo, en la figura 2.6, el ángulo del sector correspondiente al cuchillo como arma debe ser $360(0{,}18) = 64{,}8^{\circ}$.

Problemas

1. Los siguientes datos representan los tamaños de 30 familias que residen en una pequeña ciudad de Guatemala.

   5, 13, 9, 12, 7, 4, 8, 6, 6, 10, 7, 11, 10, 8, 15,
   8, 6, 9, 12, 10, 7, 11, 10, 8, 12, 9, 7, 10, 7, 8

(a) Construya una tabla de frecuencias para estos datos.

(b) Represente estos datos mediante un gráfico de líneas.

(c) Represente gráficamente los datos mediante un polígono de frecuencias.

2. La siguiente tabla de frecuencias representa las ventas semanales de bicicletas de un comercio durante un periodo de 42 semanas.

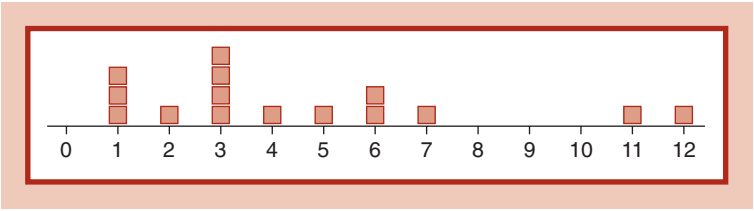
Valor	0	1	2	3	4	5	6	7
Frecuencia	3	6	7	10	8	5	2	1

(a) ¿En cuántas semanas se vendieron al menos 2 bicicletas?

(b) ¿En cuántas semanas se vendieron al menos 5 bicicletas?

(c) ¿En cuántas semanas se vendió un número par de bicicletas?

3. A 15 alumnos de cuarto curso se les preguntó a cuántas manzanas vivían de la escuela. Los resultados aparecen en el siguiente gráfico.



(a) ¿A qué número máximo de manzanas hasta la escuela se encuentra el domicilio de un alumno?

(b) ¿Cuál es el número mínimo de manzanas?

(c) ¿Cuántos alumnos viven a menos de 5 manzanas de la escuela?

(d) ¿Cuántos alumnos viven a más de 4 manzanas de la escuela?

4. Determine cuáles de los siguientes conjuntos de datos son simétricos, aproximadamente simétricos, o totalmente asimétricos.

A: 6, 0, 2, 1, 8, 3, 5
B: 4, 0, 4, 0, 2, 1, 3, 2
C: 1, 1, 0, 1, 0, 3, 3, 2, 2, 2
D: 9, 9, 1, 2, 3, 9, 8, 4, 5

5. La tabla siguiente lista los valores pero sólo algunas de sus frecuencias, para un conjunto de datos simétrico. Rellene las frecuencias que faltan.

Valor	Frecuencia
10	8
20	
30	7
40	
50	3
60	

6. Los siguientes valores representan las calificaciones de 32 estudiantes que se presentaron a un examen de Estadística.

   55, 70, 80, 75, 90, 80, 60, 100, 95, 70, 75, 85, 80, 80, 70, 95,
   100, 80, 85, 70, 85, 90, 80, 75, 85, 70, 90, 60, 80, 70, 85, 80

   Represente estos datos mediante una tabla de frecuencias, y dibuje después un gráfico de barras.

7. Dibuje una tabla de frecuencias relativas para los datos del problema 1. Dibuje estas frecuencias relativas mediante un gráfico de líneas.

8. Los siguientes datos representan los tiempos de progresión, medidos en meses, de un tipo particular de tumor cerebral, llamado *glioblastoma*, en 65 pacientes:

   6, 5, 37, 10, 22, 9, 2, 16, 3, 3, 11, 9, 5, 14, 11, 3, 1, 4, 6, 2, 7,
   3, 7, 5, 4, 8, 2, 7, 13, 16, 15, 9, 4, 4, 2, 3, 9, 5, 11, 3, 7, 5, 9,
   3, 8, 9, 4, 10, 3, 2, 7, 6, 9, 3, 5, 4, 6, 4, 14, 3, 12, 6, 8, 12, 7

   (a) Construya una tabla de frecuencias relativas para este conjunto de datos.

   (b) Represente gráficamente las frecuencias relativas mediante un polígono de frecuencias.

   (c) ¿Es este conjunto de datos aproximadamente simétrico?

9. La siguiente tabla de frecuencias relativas se ha obtenido a partir de los datos registrados sobre el número mensual de operaciones de emergencia de apendicitis que se han llevado a cabo en un determinado hospital.

Valor	0	1	2	3	4	5	6	7
Frecuencia relativa	0,05	0,08	0,12	0,14	0,16	0,20	0,15	0,10

   (a) ¿En qué proporción de meses ha habido menos de 2 operaciones de apendicitis de emergencia?

   (b) ¿En qué proporción de meses ha habido más de 5 operaciones?

   (c) ¿Es este conjunto de datos simétrico?

10. Las tablas y los gráficos de frecuencias relativas son particularmente útiles cuando se quieren comparar distintos conjuntos de datos. Los dos siguientes conjuntos de datos se refieren al número de meses que transcurrieron, en los primeros años de la epidemia, entre el diagnóstico y la muerte para dos muestras de pacientes con SIDA, varones y mujeres.

Hombres	15	13	16	10	8	20	14	19	9	12	16	18	20	12	14	14
Mujeres	8	12	10	8	14	12	13	11	9	8	9	10	14	9	10	

Represente en la misma gráfica los dos grupos de datos mediante polígonos de frecuencias. Utilice un color diferente para cada conjunto de datos. ¿Qué conclusión se puede sacar respecto a qué conjunto de datos que tiende a tener valores mayores?

11. Con los datos del ejemplo 2.2, determine la proporción de puntuaciones ganadoras del Torneo de Maestros de Golf que:

    (a) Son inferiores a 280

    (b) Son iguales o superiores a 282.

    (c) Están comprendidas entre 278 y 284, ambos inclusive

La tabla siguiente muestra el número medio de días de cada mes con al menos 0,01 pulgadas de lluvia en varias ciudades. Los problemas del 12 al 14 se refieren a ella.

Número medio de días con una precipitación de 0,01 pulgadas o más

Estado	Ciudad	Longitud de registro	Ene.	Feb.	Mar.	Abr.	May.	Jun.	Jul.	Ago.	Sep.	Oct.	Nov.	Dic.	Anual
AL	Mobile	46	11	10	11	7	8	11	16	14	10	6	8	10	123
AK	Juneau	43	18	17	18	17	17	16	17	18	20	24	19	21	220
AZ	Phoenix	48	4	4	4	2	1	1	4	5	3	3	3	4	36
AR	Little Rock	45	9	9	10	10	10	8	8	7	7	7	8	9	103
CA	Los Angeles	52	6	6	6	3	1	1	1	0	1	2	4	5	36
	Sacramento	48	10	9	9	5	3	1	0	0	1	3	7	9	58
	San Diego	47	7	6	7	5	2	1	0	1	1	3	5	6	43
	San Francisco	60	11	10	10	6	3	1	0	0	1	4	7	10	62
CO	Denver	53	6	6	9	9	11	9	9	9	6	5	5	5	89
CT	Hartford	33	11	10	11	11	12	11	10	10	9	8	11	12	127
DE	Wilmington	40	11	10	11	11	11	10	9	9	8	8	10	10	117
DC	Washington	46	10	9	11	10	11	10	10	9	8	7	8	9	111
FL	Jacksonville	46	8	8	8	6	8	12	15	14	13	9	6	8	116
	Miami	45	6	6	6	6	10	15	16	17	17	14	9	7	129
GA	Atlanta	53	11	10	11	9	9	10	12	9	8	6	8	10	115
HI	Honolulu	38	10	9	9	9	7	6	8	6	7	9	9	10	100
ID	Boise	48	12	10	10	8	8	6	2	3	4	6	10	11	91
IL	Chicago	29	11	10	12	12	11	10	10	9	10	9	10	12	127
	Peoria	48	9	8	11	12	11	10	9	8	9	8	9	10	114
IN	Indianapolis	48	12	10	13	12	12	10	9	9	8	8	10	12	125
IA	Des Moines	48	7	7	10	11	11	11	9	9	9	8	7	8	107
KS	Wichita	34	6	5	8	8	11	9	7	8	8	6	5	6	86
KY	Louisville	40	11	11	13	12	12	10	11	8	8	8	10	11	125
LA	New Orleans	39	10	9	9	7	8	11	15	13	10	6	7	10	114
ME	Portland	47	11	10	11	12	13	11	10	9	8	9	12	12	128
MD	Baltimore	37	10	9	11	11	11	9	9	10	7	7	9	9	113

(Continuación)

Número medio de días con una precipitación de 0,01 pulgadas o más *(Continuación)*

Estado	Ciudad	Longitud de registro	Ene.	Feb.	Mar.	Abr.	May.	Jun.	Jul.	Ago.	Sep.	Oct.	Nov.	Dic.	Anual
MA	Boston	36	12	10	12	11	12	11	9	10	9	9	11	12	126
MI	Detroit	29	13	11	13	12	11	11	9	9	10	9	12	14	135
	Sault Ste. Marie	46	19	15	13	11	11	12	10	11	13	13	17	20	165
MN	Duluth	46	12	10	11	10	12	13	11	11	12	10	11	12	134
	Minneapolis-St. Paul	49	9	7	10	10	11	12	10	10	10	8	8	9	115
MS	Jackson	24	11	9	10	8	10	8	10	10	8	6	8	10	109
MO	Kansas City	15	7	7	11	11	11	11	7	9	8	8	8	8	107
	St. Louis	30	8	8	11	11	11	10	8	8	8	8	10	9	111
MT	Great Falls	50	9	8	9	9	12	12	7	8	7	6	7	8	101
NE	Omaha	51	6	7	9	10	12	11	9	9	9	7	5	6	98
NV	Reno	45	6	6	6	4	4	3	2	2	2	3	5	6	51
NH	Concord	46	11	10	11	12	12	11	10	10	9	9	11	11	125
NJ	Atlantic City	44	11	10	11	11	10	9	9	9	8	7	9	10	112
NM	Albuquerque	48	4	4	5	3	4	4	9	9	6	5	3	4	61
NY	Albany	41	12	10	12	12	13	11	10	10	10	9	12	12	134
	Buffalo	44	20	17	16	14	12	10	10	11	11	12	16	20	169
	New York	118	11	10	11	11	11	10	10	10	8	8	9	10	121
NC	Charlotte	48	10	10	11	9	10	10	11	9	7	7	8	10	111
	Raleigh	43	10	10	10	9	10	9	11	10	8	7	8	9	111
ND	Bismarck	48	8	7	8	8	10	12	9	9	7	6	6	8	97
OH	Cincinnati	40	12	11	13	13	11	11	10	9	8	8	11	12	129
	Cleveland	46	16	14	15	14	13	11	10	10	10	11	14	16	156
	Columbus	48	13	12	14	13	13	11	11	9	8	9	11	13	137
OK	Oklahoma City	48	5	6	7	8	10	9	6	6	7	6	5	5	82
OR	Portland	47	18	16	17	14	12	9	4	5	8	13	18	19	152
PA	Philadelphia	47	11	9	11	11	11	10	9	9	8	8	9	10	117
	Pittsburgh	35	16	14	16	14	12	12	11	10	9	11	13	17	154
RI	Providence	34	11	10	12	11	11	11	9	10	8	8	11	12	124
SC	Columbia	40	10	10	11	8	9	9	12	11	8	6	7	9	109
SD	Sioux Falls	42	6	6	9	9	10	11	9	9	8	6	6	6	97
TN	Memphis	37	10	9	11	10	9	8	9	8	7	6	9	10	106
	Nashville	46	11	11	12	11	11	9	10	9	8	7	10	11	119
TX	Dallas-Fort Worth	34	7	7	7	8	9	6	5	5	7	6	6	6	78
	El Paso	48	4	3	2	2	2	4	8	8	5	4	3	4	48
	Houston	18	10	8	9	7	9	9	9	10	10	8	9	9	106
UT	Salt Lake City	59	10	9	10	9	8	5	5	6	5	6	8	9	91
VT	Burlington	44	14	12	13	12	14	13	12	12	12	12	14	15	154
VA	Norfolk	39	10	10	11	10	10	9	11	10	8	8	8	9	114
	Richmond	50	10	9	11	9	11	9	11	10	8	7	8	9	113
WA	Seattle	43	19	16	17	14	10	9	5	6	9	13	18	20	156
	Spokane	40	14	12	11	9	9	8	4	5	6	8	12	15	113
WV	Charleston	40	16	14	15	14	13	11	13	11	9	10	12	14	151
WI	Milwaukee	47	11	10	12	12	12	11	10	9	9	9	10	11	125
WY	Cheyenne	52	6	6	9	10	12	11	11	10	7	6	6	5	99
PR	San Juan	32	16	13	12	13	17	16	19	18	17	17	18	19	195

Fuente: Administración Atmosférica y Oceánica de Estados Unidos, *Datos climáticos comparativos*.

12. Construya una tabla de frecuencias relativas para el número medio de días de lluvia en enero en las diferentes ciudades. A continuación, represente gráficamente los datos mediante un polígono de frecuencias relativas.
13. Usando solamente los datos relativos a las 12 primeras ciudades de la lista, construya una tabla de frecuencias para el número medio de días de lluvia en los meses de noviembre y diciembre.
14. Usando sólo los datos que se refieren a las 24 primeras ciudades, construya una tabla de frecuencias relativas para el mes de junio y, por separado, otra para el mes de diciembre. Posteriormente, represente en un mismo gráfico los dos conjuntos de datos mediante polígonos de frecuencias relativas.
15. La tabla siguiente muestra el número de muertes que hubo en las carreteras británicas durante 1987 distribuidas por clases.

Clases	**Número de muertes**
Peatones	1699
Ciclistas	280
Motoristas	650
Conductores de automóviles	1327

Represente estos datos mediante un gráfico de tarta.

16. Los siguientes datos, sacados del *New York Times*, representan los porcentajes, por kilos de peso, de los distintos componentes de la basura de Nueva York. Represéntelos mediante un gráfico de tarta.

Materia orgánica (comida, desperdicios del jardín, madera, etc.)	37,3
Papel	30,8
Bultos (mobiliario, neveras, etc.)	10,9
Plástico	8,5
Cristal	5
Metal	4
Inorgánicos	2,2
Aluminio	0,9
Desperdicios peligrosos	0,4

2.3 Datos agrupados e histogramas

Como se ha visto en la sección 2.2, el uso de gráficos de barras o líneas es una forma bastante efectiva de representar las frecuencias de los diferentes valores. Sin embargo, en algunos conjuntos de datos el número de valores distintos es demasiado grande para que se puedan utilizar los gráficos citados. En su lugar, es posible clasificar dichos valores en grupos, o *intervalos de clase*, para luego representar gráficamente el número de datos que corresponden a cada clase. En la elección del número de intervalos de clase se debe ponderar entre: (i) elegir pocos a costa de perder mucha información sobre los datos reales de cada intervalo de clase, o (ii) elegir muchos, con lo que las frecuencias resultantes de cada

intervalo de clase pueden ser demasiado pequeñas para que se reconozcan los patrones de forma. Aunque lo más habitual suele ser entre 5 y 10 intervalos de clase, el número apropiado es una elección subjetiva, y uno puede, como es natural, probar distintos números de intervalos de clase para ver cuál de los gráficos resultantes revela más información sobre los datos. Es corriente, aunque no esencial, elegir intervalos de clase de igual longitud.

Los puntos inicial y final de cada intervalo de clase se llaman extremos del mismo. Nosotros utilizaremos el convenio de *inclusión por la izquierda*, lo que significa que el intervalo de clase incluye el extremo de la izquierda pero no el de la derecha. Por ejemplo, el intervalo 20-30 incluye los valores que son mayores *o iguales* que 20 y menores que 30.

Los datos de la tabla 2.5 representan los niveles de colesterol en la sangre de 40 estudiantes de primer curso de una cierta universidad. Antes de determinar las clases y sus frecuencias, es útil ordenar los datos de forma creciente, así se consiguen los 40 valores de la tabla 2.6.

Puesto que los datos varían entre el valor mínimo (171) y el máximo (227), el extremo de la izquierda de la primera clase debe ser menor o igual a 171, y el extremo de la derecha de la última clase debe ser mayor que 227. Podría elegirse tomar como primera clase el intervalo de 170 a 180, lo que nos lleva a tomar seis clases. La tabla 2.7 nos muestra las frecuencias (y también las frecuencias relativas) de los valores de datos que caen dentro de cada intervalo de clase.

Observación: Debido al convenio de inclusión por la izquierda, los valores iguales a 200 se colocarán dentro del intervalo con extremos 200 y 210, y no en el intervalo comprendido entre 190 y 200.

Un gráfico de barras en el que las barras sean adyacentes se llama *histograma*. El eje vertical de un histograma puede representar bien las frecuencias de los intervalos de clase o bien sus frecuencias relativas. En el primer caso, el histograma se llama *histograma de frecuencias*; en el segundo, se trata de un *histograma de frecuencias relativas*. La figura 2.7 presenta un histograma de frecuencias para los datos de la tabla 2.7.

Es importante saber que una tabla de frecuencias de intervalos de clase o un histograma basado en tal tabla no contiene toda la información del conjunto de datos originales. Ambas representaciones utilizan sólo el número de valores dentro de cada intervalo de clase, y no los valores reales de los datos. Así pues, aunque las tablas y los gráficos citados son un útil reflejo de los datos, el conjunto de datos originales se debe mantener *siempre*.

Tabla 2.5 Niveles de colesterol en la sangre

213	174	193	196	220	183	194	200
192	200	200	199	178	183	188	193
187	181	193	205	196	211	202	213
216	206	195	191	171	194	184	191
221	212	221	204	204	191	183	227

Tabla 2.6 Niveles de colesterol en la sangre en orden creciente

171, 174, 178, 181, 183, 183, 183, 184,187, 188, 191, 191, 191, 192, 193, 193, 193, 194, 194, 195, 196, 196, 199, 200, 200, 200, 202, 204, 204, 205, 206, 211, 212, 213, 213, 216, 220, 221, 221, 227

Tabla 2.7 Tabla de frecuencias de los niveles de colesterol en la sangre

Intervalos de clase	Frecuencias	Frecuencias relativas
170–180	3	$\frac{3}{40} = 0{,}075$
180–190	7	$\frac{7}{40} = 0{,}175$
190–200	13	$\frac{13}{40} = 0{,}325$
200–210	8	$\frac{8}{40} = 0{,}20$
210–220	5	$\frac{5}{40} = 0{,}125$
220–230	4	$\frac{4}{40} = 0{,}10$

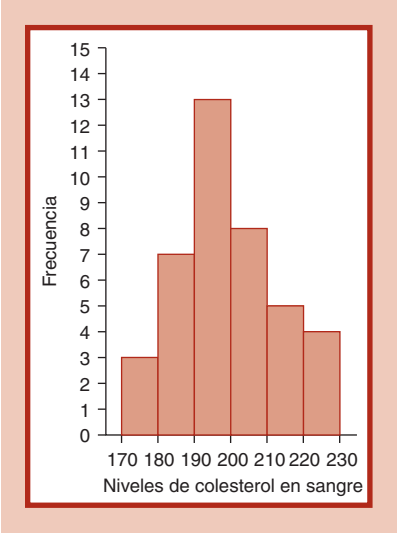


Figura 2.7 Histograma de frecuencias para los datos de la tabla 2.7.

Para construir un histograma a partir de un conjunto de datos

1. Ordene los datos en forma creciente.
2. Elija los intervalos de clase de manera que todos los datos aparezcan en alguno de ellos.
3. Construya una tabla de frecuencias.
4. Dibuje las barras adyacentes con alturas iguales a las frecuencias del paso 3.

La importancia de un histograma estriba en que permite organizar y presentar los datos gráficamente para que se pueda prestar atención a determinadas características importantes de los datos. Por ejemplo, un histograma puede a menudo indicar:

1. La simetría de los datos.
2. La dispersión de éstos.
3. Si existen intervalos que tienen un alto nivel de concentración de datos.
4. Si existen brechas entre los datos.
5. Si algunos valores de datos están muy separados de otros.

Por ejemplo, el histograma presentado en la figura 2.7 indica que las frecuencias de las sucesivas clases primero crecen y luego decrecen, y alcanzan un máximo en el intervalo de clase comprendido entre 190 y 200. Los histogramas de la figura 2.8 proporcionan una

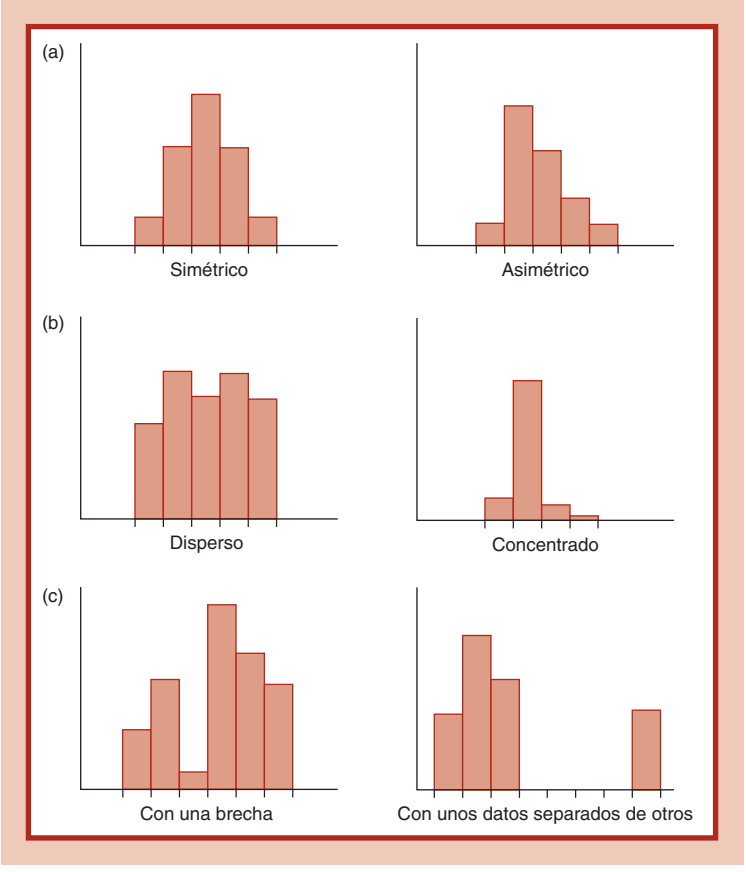


Figura 2.8 Características de los datos detectadas por los histogramas: (a) simetría, (b) grado de dispersión y dónde se concentran los valores, y (c) brechas en los datos y datos muy separados de otros.

información valiosa sobre los conjuntos de datos que representan. El conjunto de datos cuyo histograma se encuentra a la izquierda de la figura 2.8(a) es simétrico, mientras que el que se representa a la derecha no lo es. El conjunto de datos representado a la izquierda de la figura 2.8(b) se encuentra bastante disperso, mientras que el que se muestra a la derecha está más concentrado. El conjunto de datos representado a la izquierda de la figura 2.8(c) presenta una brecha, mientras que el representado al lado derecho tiene ciertos valores alejados del resto.

Ejemplo 2.3 La tabla 2.8 muestra las tasas de natalidad (por 1000 habitantes) en cada uno de los Estados de Estados Unidos. Represente gráficamente estos datos en un histograma.

Solucón Puesto que los datos varían entre el valor más bajo, (12,4), y el más alto, (21,9), usaremos intervalos de clase de longitud 1,5, comenzando en el valor 12. Con estos intervalos de clase se obtiene la siguiente tabla de frecuencias.

Intervalos de clase	Frecuencias	Intervalos de clase	Frecuencias
12,0–13,5	2	18,0–19,5	2
13,5–15,0	15	19,5–21,0	0
15,0–16,5	22	21,0–22,5	2
16,5–18,0	7		

Un gráfico de histograma para estos datos se presenta en la figura 2.9.

Tabla 2.8 Tasas de natalidad por cada 100 habitantes

Estado	Tasa	Estado	Tasa	Estado	Tasa
Alabama	14,2	Louisiana	15,7	Ohio	14,9
Alaska	21,9	Maine	13,8	Oklahoma	14,4
Arizona	19,0	Maryland	14,4	Oregon	15,5
Arkansas	14,5	Massachusetts	16,3	Pennsylvania	14,1
California	19,2	Michigan	15,4	Rhode Island	15,3
Colorado	15,9	Minnesota	15,3	South Carolina	15,7
Connecticut	14,7	Mississippi	16,1	South Dakota	15,4
Delaware	17,1	Missouri	15,5	Tennessee	15,5
Florida	15,2	Montana	14,1	Texas	17,7
Georgia	17,1	Nebraska	15,1	Utah	21,2
Hawaii	17,6	Nevada	16,5	Vermont	14,0
Idaho	15,2	New Hampshire	16,2	Virginia	15,3
Illinois	16,0	New Jersey	15,1	Washington	15,4
Indiana	14,8	New Mexico	17,9	West Virginia	12,4
Iowa	13,1	New York	16,2	Wisconsin	14,8
Kansas	14,2	North Carolina	15,6	Wyoming	13,7
Kentucky	14,1	North Dakota	16,5		

Fuente: Departamento de Salud y Servicios Sociales.

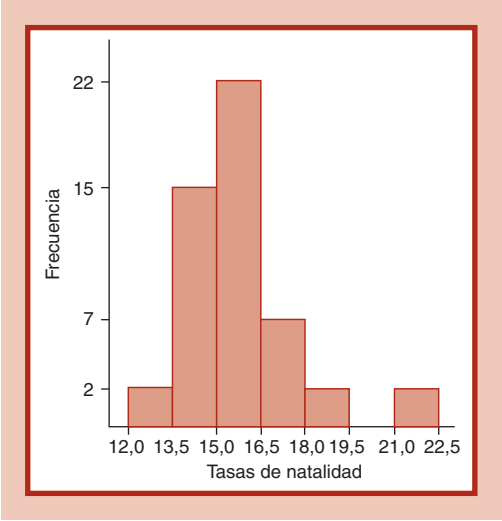


Figura 2.9 Histograma para las tasas de natalidad de los 50 Estados.

Un histograma es, en esencia, un diagrama de barras que muestra gráficamente las frecuencias o las frecuencias relativas de los datos que aparecen dentro de los distintos intervalos de clase. Dichas frecuencias de clase también se pueden representar gráficamente mediante polígonos de frecuencias (o de frecuencias relativas). Cada intervalo de clase es identificado por un valor, que generalmente coincide con el punto medio del intervalo. Después, estos valores se representan gráficamente frente a las frecuencias de los intervalos de clase que representan y los puntos del gráfico se conectan mediante líneas rectas para conseguir el polígono de frecuencias. Estos gráficos son especialmente útiles para comparar conjuntos de datos, puesto que en un mismo gráfico se pueden mostrar varios polígonos de frecuencias. ■

Ejemplo 2.4 Los datos de la tabla 2.9 representan las frecuencias de los intervalos de clase para las presiones sistólicas sanguíneas de dos grupos de trabajadores industriales: aquellos cuya edad está comprendida entre 30 y 40 años, y aquellos cuya edad se encuentra entre 50 y 60 años.

Resulta difícil comparar directamente las presiones sanguíneas de los dos grupos de edad dado que el número total de trabajadores de cada grupo es diferente. Para salvar esta dificultad, se pueden computar y representar gráficamente las frecuencias *relativas* de cada una de las clases. Es decir, todas las frecuencias referidas a los trabajadores cuya edad varía entre 30 y 39 años se dividen entre 2540 (el número de dichos trabajadores) y todas las frecuencias referidas a los trabajadores con edades entre 50 y 59 años se dividen entre 731. Los resultados se muestran en la tabla 2.10.

La figura 2.10 es un gráfico de los polígonos de frecuencias relativas para ambos grupos de edad. Si se visualizan ambos polígonos de frecuencia en un mismo gráfico resulta fácil comparar los dos conjuntos de datos. Por ejemplo, aparentemente las presiones sanguíneas del grupo de mayor edad tienden a extenderse sobre valores más altos que los del grupo más joven.

Tabla 2.9 Frecuencias de clase de la presión sanguínea sistólica de dos grupos de trabajadores varones.

	Número de trabajadores	
Presión sanguínea	**Edad entre 30-40 años**	**Edad entre 50-60 años**
Menos de 90	3	1
90–100	17	2
100–110	118	23
110–120	460	57
120–130	768	122
130–140	675	149
140–150	312	167
150–160	120	73
160–170	45	62
170–180	18	35
180–190	3	20
190–200	1	9
200–210		3
210–220		5
220–230		2
230–240		1
Total	2540	731

Tabla 2.10 Frecuencias relativas de clase para las presiones sanguíneas.

	Porcentaje de trabajadores	
Presión sanguínea	**Edad entre 30-40 años**	**Edad entre 50-60 años**
Menos de 90	0,12	0,14
90–100	0,67	0,27
100–110	4,65	3,15
110–120	18,11	7,80
120–130	30,24	16,69
130–140	26,57	20,38
140–150	12,28	22,84
150–160	4,72	9,99
160–170	1,77	8,48
170–180	0,71	4,79
180–190	0,12	2,74
190–200	0,04	1,23
200–210		0,41
210–220		0,68
220–230		0,27
230–240		0,14
Total	100,00	100,00

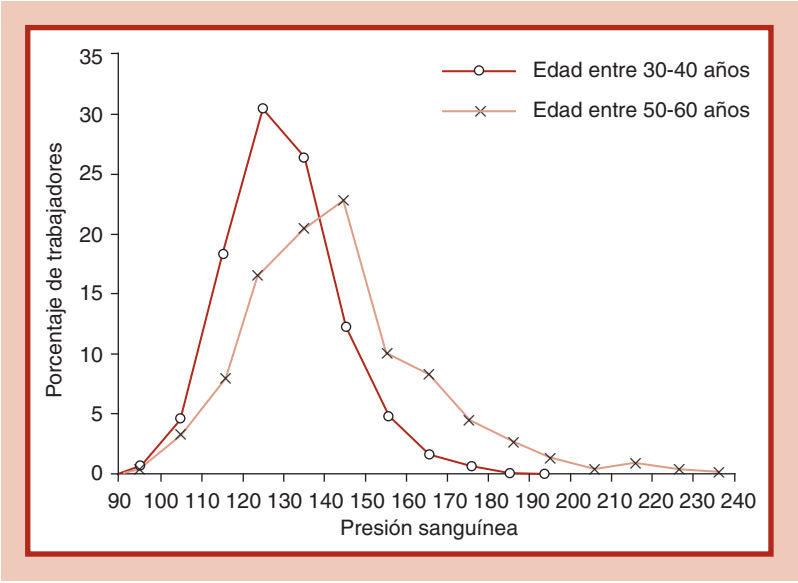


Figura 2.10 Polígonos de frecuencias relativas.

Problemas

1. Los siguientes conjuntos de datos representan las puntuaciones obtenidas por 40 estudiantes de sexto curso en un test de cociente de inteligencia (IQ, *Intelligence Quotient*) en una determinada escuela:

   114, 122, 103, 118, 99, 105, 134, 125, 117, 106, 109, 104, 111, 127,
   133, 111, 117, 103, 120, 98, 100, 130, 141, 119, 128, 106, 109, 115,
   113, 121, 100, 130, 125, 117, 119, 113, 104, 108, 110, 102

   (a) Presente este conjunto de datos en un histograma de frecuencias.

   (b) ¿Qué intervalo de clase contiene el mayor número de valores de datos?

   (c) ¿Existe, grosso modo, el mismo número de datos en cada uno de los intervalos de clase?

   (d) ¿El histograma parece aproximadamente simétrico?

2. Los siguientes datos muestran las temperaturas máximas (en grados Celsius), de los días 4 de julio de 30 años sucesivos, en la ciudad de San Francisco:

   22,8, 26,2, 31,7, 31,1, 26,9, 28,0, 29,4, 28,8, 26,7, 27,4, 28,2,
   30,3, 29,5, 28,9, 27,5, 28,3, 24,1, 25,3, 28,5, 27,7, 24,4,
   29,2, 30,3, 33,7, 27,5, 29,3, 30,2, 28,5, 32,2, 33,7

   (a) Presente este conjunto de datos en un histograma de frecuencias.

   (b) ¿Cuál diría que es la temperatura máxima "típica" de un 4 de julio en San Francisco?

   (c) ¿Qué otras conclusiones pueden extraerse del histograma?

3. Los siguientes datos (en miles de dólares) representan las rentas netas anuales de una muestra de contribuyentes:

   47, 55, 18, 24, 27, 41, 50, 38, 33, 29, 15, 77, 64, 22, 19, 35, 39, 41,
   67, 55, 121, 77, 80, 34, 41, 48, 60, 30, 22, 28, 84, 55, 26, 105, 62,
   30, 17, 23, 31, 28, 56, 64, 88, 104, 115, 39, 25, 18, 21, 30, 57, 40,
   38, 29, 19, 46, 40, 49, 72, 70, 37, 39, 18, 22, 29, 52, 94, 86, 23, 36

   (a) Represente gráficamente estos datos mediante un histograma de frecuencias con 5 intervalos de clase.

   (b) Represente gráficamente estos datos mediante un histograma de frecuencias con 10 intervalos de clase.

   (c) ¿Cuál de los dos histogramas cree que es más informativo? ¿Por qué?

4. Un conjunto de 200 datos puntuales se dividió en 8 clases, todas de tamaño 3 (en las unidades de los datos). Después se determinaron las frecuencias de cada clase y se construyó una tabla de frecuencias. Sin embargo, ciertas entradas de esta tabla se perdieron. Supongamos que la parte de la tabla de frecuencias que se conservó es la siguiente:

Intervalo de clase	Frecuencia	Frecuencia relativa
		0,05
	14	
	18	
15–18	38	
		0,10
	42	
	11	

   Rellene los valores perdidos de la tabla y dibuje un histograma de frecuencias relativas.

5. Los siguientes valores muestran las concentraciones de ozono (medidas en partes por 100 millones) en el aire del centro de la ciudad de Los Ángeles durante 25 días consecutivos del verano de 1984:

   6,2, 9,1, 2,4, 3,6, 1,9, 1,7, 4,5, 4,2, 3,3, 5,1, 6,0, 1,8, 2,3,
   4,9, 3,7, 3,8, 5,5, 6,4, 8,6, 9,3, 7,7, 5,4, 7,2, 4,9, 6,2

   (a) Construya un histograma de frecuencias para este conjunto de datos, uno de cuyos intervalos de clase vaya de 3 a 5.

   (b) Construya un histograma de frecuencias para este conjunto de datos, uno de cuyos intervalos de clase vaya de 2 a 3.

   (c) ¿Qué histograma de frecuencias crees que es más informativo?

6. Los siguientes datos reflejan las producciones de carne, en miles de toneladas métricas, del año 2002 en 11 países distintos:

País	Producción	País	Producción
Argentina	2,748	Japón	520
Australia	2,034	México	1,450
Brasil	7,150	España	592
China	5,616	Reino Unido	1,390
Francia	1,666	Estados Unidos	12,424
Italia	1,161		

   (a) Represente estos datos en un histograma de frecuencias.

   (b) Un valor de dato que se encuentra muy separado del resto se denomina un *outlier*, o valor extremo. ¿Existe algún *outlier* en el conjunto de datos dado?

7. Considere los niveles de colesterol en la sangre de los primeros 100 estudiantes que aparecen en el conjunto del Apéndice A. Divida estos estudiantes en dos grupos por sexo, y construya una tabla de frecuencias relativas para cada uno de ellos. Dibuje en un mismo gráfico los polígonos de frecuencias relativas para los estudiantes varones y hembras. ¿Se pueden extraer conclusiones sobre la relación existente entre el sexo y el nivel de colesterol?

8. Utilice la siguiente tabla para construir un histograma de frecuencias de las cantidades que, por cada dólar recaudado por impuestos, el gobierno federal devuelve a los diferentes Estados.

Devolución federal a los Estados por cada dólar recaudado mediante impuestos (cifras del año fiscal 2002, ordenadas de mayor a menor)

Estado	Devolución	Estado	Devolución	Estado	Devolución	Estado	Devolución	Estado	Devolución
District of Columbia	$6,44	Arkansas	$1,55	Maryland	$1,22	Ohio	$1,03	Delaware	$0,85
		Oklahoma	1,52	Arizona	1,21	Georgia	1,01	Colorado	0,78
New Mexico	2,37	Virginia	1,50	Nebraska	1,19	Florida	1,01	Minnesota	0,77
North Dakota	2,07	Kentucky	1,50	Utah	1,14	Indiana	1,00	Illinois	0,77
Alaska	1,91	Louisiana	1,48	Kansas	1,13	Oregon	0,98	California	0,76
Mississippi	1,89	South Carolina	1,34	Vermont	1,13	Texas	0,92	Massachusetts	0,75
West Virginia	1,82	Maine	1,34	Pennsylvania	1,09	Wisconsin	0,88	Nevada	0,74
Montana	1,67	Missouri	1,34	Rhode Island	1,08	Michigan	0,88	New Hampshire	0,66
Alabama	1,64	Idaho	1,31	North Carolina	1,07	Washington	0,87	Connecticut	0,65
South Dakota	1,61	Tennessee	1,26	Wyoming	1,06	New York	0,85	New Jersey	0,62
Hawaii	1,57	Iowa	1,23						

Fuente: Administración Fiscal.

La tabla siguiente muestra los datos relativos a las tasas de mortalidad accidental en Estados Unidos durante varios años. Utilícela para contestar a los problemas de 9 a 12.

Tasas de defunción por cada 100 000 habitantes para los principales tipos de muerte accidental en los Estados Unidos, 1970-2002

Año	Vehículos a motor	Caídas	Envene-namientos	Ahogamientos	Fuegos, incendios, humo	Ingestión de comida u objetos	Armas de fuego
1970	26,8	8,3	2,6	3,9	3,3	1,4	1,2
1980	23,4	5,9	1,9	3,2	2,6	1,4	0,9
1985	19,3	5,0	2,2	2,2	2,1	1,5	0,7
1990	18,8	4,9	2,3	1,9	1,7	1,3	0,6
1991	17,3	5,0	2,6	1,8	1,6	1,3	0,6
1992	16,1	5,0	2,7	1,4	1,6	1,2	0,6
1993	16,3	5,1	3,4	1,5	1,5	1,2	0,6
1994	16,3	5,2	3,5	1,5	1,5	1,2	0,5
1995	16,5	5,3	3,4	1,7	1,4	1,2	0,5
1996	16,5	5,6	3,5	1,5	1,4	1,2	0,4
1997	16,2	5,8	3,8	1,5	1,3	1,2	0,4
1998	16,1	6,0	4,0	1,6	1,2	1,3	0,3
1999	15,5	4,8	4,5	1,3	1,2	1,4	0,3
2000	15,7	4,8	4,6	1,3	1,2	1,6	0,3
2001	15,7	5,1	5,0	1,2	1,2	1,4	0,3
2002	15,7	5,2	5,6	1,1	1,0	1,5	0,3

Fuente: Consejo de Seguridad Nacional.

9. Construya un histograma de frecuencias relativas para las tasas de mortalidad anuales producidas con vehículos a motor.

10. Construya un histograma de frecuencias relativas para las tasas de mortalidad anuales debidas a caídas.

11. Construya un histograma de frecuencias relativas para las tasas anuales de mortalidad para el total de causas citadas.

12. ¿Diría que las tasas de mortalidad accidental se mantienen relativamente constantes?

13. A partir de la tabla que antecede al problema 12 de la sección 2.2, construya un histograma del número medio de días de lluvia en las ciudades de la lista.

14. Considere la tabla siguiente.

Edad del conductor, en años	Porcentaje de conductores	Porcentaje de conductores con accidentes fatales
15–20	9	18
20–25	13	21
25–30	13	14
30–35	11	11
35–40	9	7
40–45	8	6
45–50	8	5
50–55	7	5
55–60	6	4
60–65	6	3
65–70	4	2
70–75	3	2
Más de 75	3	2

Por el criterio de inclusión por la izquierda de las clases, el 13% del total de los conductores tienen como mínimo 25 años y menos de 30, y un 11% de los conductores muertos en accidentes de coche tienen por lo menos 30 años y menos de 35.

(a) Dibuje un histograma de frecuencias relativas para las clases de edad de los conductores.

(b) Dibuje un histograma de frecuencias relativas para las clases de edad de los conductores muertos en accidentes de coche.

(c) ¿Cuál es el grupo que tiene un mayor número de accidentes fatales?

(d) ¿Cuál es el grupo que merecería mayores descuentos en los seguros? Explique su razonamiento.

15. Las tablas de frecuencias relativas acumuladas muestran el porcentaje de valores de datos menores que un valor dado, para una sucesión creciente de valores. Dichas tablas se pueden construir a partir de tablas de frecuencias relativas mediante la suma de las frecuencias relativas de forma acumulada. La tabla siguiente muestra los valores iniciales de la tabla acumulada citada para los dos conjuntos de datos incluidos en la tabla 2.9. Muestra, por ejemplo, que el 5,44% de los hombres con unas edades comprendidas entre 30 y 40 años tiene presiones sanguíneas inferiores a 110, y que sólo un 3,56% de los que tienen entre 50 y 60 años tienen una presión inferior a la citada.

Tabla de frecuencias relativas acumuladas para los conjuntos de datos de la tabla 2.9

	Porcentaje de trabajadores	
Presión sanguínea menor de	**Con edad de 30-40**	**Con edad de 50-60**
90	0,12	0,14
100	0,79	0,41
110	5,44	3,56
120		
130		
.		
.		
.		
240	100	100

(a) Explique por qué la frecuencia relativa acumulada de la última clase debe ser 100.

(b) Complete la tabla.

(c) ¿Qué indica la tabla sobre los dos conjuntos de datos? (Esto es, ¿cuál tiende a tener valores menores?)

(d) Represente en un mismo gráfico los polígonos de frecuencias relativas acumuladas de los datos dados. Tales gráficos se denominan *ojivas*.

2.4 Gráficos de tallos y hojas

Una forma eficiente de representar un conjunto de datos de tamaño pequeño o moderado consiste en utilizar los *gráficos de tallos y hojas*. Tales gráficos se obtienen dividiendo cada valor de dato en dos partes –su tallo y su hoja–. Por ejemplo, si todos los datos son de dos dígitos, el tallo de un valor podría ser el dígito de las decenas y su hoja, el dígito de las unidades. Es decir, el valor 84 se expresaría como

Tallo	Hoja
8	4

y los dos valores 84 y 87 se representarían conjuntamente de la siguiente manera

Tallo	Hoja
8	4, 7

Ejemplo 2.5 La tabla 2.11 presenta las rentas per cápita para los 50 Estados de Estados Unidos y para el Distrito de Columbia.

Tabla 2.11 Rentas per cápita (en dólares por persona), en 2002

Estado		Estado	
Estados Unidos	30 941	Missouri	28 936
Alabama	25 128	Montana	25 020
Alaska	32 151	Nebraska	29 771
Arizona	26 183	Nevada	30 180
Arkansas	23 512	New Hampshire	34 334
California	32 996	New Jersey	39 453
Colorado	33 276	New Mexico	23 941
Connecticut	42 706	New York	36 043
Delaware	32 779	North Carolina	27 711
District of Columbia	42 120	North Dakota	26 982
Florida	29 596	Ohio	29 405
Georgia	28 821	Oklahoma	25 575
Hawaii	30 001	Oregon	28 731
Idaho	25 057	Pennsylvania	31 727
Illinois	33 404	Rhode Island	31 319
Indiana	28 240	South Carolina	25 400
Iowa	28 280	South Dakota	26 894
Kansas	29 141	Tennessee	27 671
Kentucky	25 579	Texas	28 551
Louisiana	25 446	Utah	24 306
Maine	27 744	Vermont	29 567
Maryland	36 298	Virginia	32 922
Massachusetts	39 244	Washington	32 677
Michigan	30 296	West Virginia	23 688
Minnesota	34 071	Wisconsin	29 923
Mississippi	22 372	Wyoming	30 578

Los datos que se muestran en la tabla 2.11 se pueden representar mediante el siguiente gráfico de tallos y hojas. Observe que los valores de las hojas aparecen en el gráfico en orden creciente.

22	372
23	512, 688, 941
24	706
25	020, 057, 128, 400, 446, 575, 579
26	183, 894, 982
27	671, 711, 744
28	240, 280, 551, 731, 821, 936
29	141, 405, 567, 596, 771, 923
30	001, 180, 296, 578
31	319, 727
32	151, 677, 779, 922, 996

33	276, 404
34	071, 334
36	043, 298
39	244, 453
42	120, 706

La elección de los tallos siempre se debe hacer de modo que el diagrama de tallos y hojas proporcione información sobre los datos. Como modelo, considere el ejemplo 2.6.

Ejemplo 2.6 Los siguientes datos representan la proporción de estudiantes de las escuelas públicas de primaria en 18 ciudades distintas.

55,2, 47,8, 44,6, 64,2, 61,4, 36,6, 28,2, 57,4, 41,3,
44,6, 55,2, 39,6, 40,9, 52,2, 63,3, 34,5, 30,8, 45,3

Si se hace que el tallo identifique el dígito de las decenas y que la hoja incluya las cifras restantes de cada valor, el gráfico de tallos y hojas resultante para los datos dados es el siguiente:

2	8,2
3	0,8, 4,5, 6,6, 9,6
4	0,9, 1,3, 4,6, 4,6, 5,3, 7,8
5	2,2, 5,2, 5,2, 7,4
6	1,4, 3,3, 4,2

Se podría haber elegido que, para cada valor, el tallo viniera representado por su parte entera y la hoja por su parte decimal, de modo que el valor 28,2 apareciera como

28 | ,2

Sin embargo, esta elección produciría demasiados tallos (con muy pocas hojas cada uno), con lo que el conjunto de datos no quedaría representado con claridad. ■

Ejemplo 2.7 Los siguientes gráficos de tallos y hojas muestran los pesos de 80 asistentes a una convención de deportes. Los tallos representan las cifras de las decenas y las hojas las cifras de las unidades.

10	2, 3, 3, 4, 7	(5)
11	0, 1, 2, 2, 3, 6, 9	(7)
12	1, 2, 4, 4, 6, 6, 6, 7, 9	(9)
13	1, 2, 2, 5, 5, 6, 6, 8, 9	(9)
14	0, 4, 6, 7, 7, 9, 9	(7)
15	1, 1, 5, 6, 6, 6, 7	(7)

16	0, 1, 1, 1, 2, 4, 5, 6, 8, 8	(10)
17	1, 1, 3, 5, 6, 6, 6	(7)
18	1, 2, 2, 5, 5, 6, 6, 9	(8)
19	0, 0, 1, 2, 4, 5	(6)
20	9, 9	(2)
21	7	(1)
22	1	(1)
23		(0)
24	9	(1)

Los números que están entre paréntesis a la derecha representan la cantidad de valores que aparecen en cada clase de tallo; son valores que habitualmente resultan útiles. Nos indican, por ejemplo, que existen 10 valores dentro del tallo 16; esto es, hay 10 individuos cuyos pesos oscilan entre 160 y 169. Observe que un tallo sin hojas (tal como el tallo con valor 23) indica que no existen ocurrencias en dicha clase.

De este gráfico resulta evidente que casi todos los valores de datos se encuentran entre 100 y 200, y que su dispersión es bastante uniforme dentro de esta región, a excepción de los escasos valores que caen dentro de los intervalos con extremos 100 y 110, y 190 y 200. ■

Los gráficos de tallos y hojas son bastante útiles para mostrar todos los valores de datos mediante una representación clara que puede ser un primer paso en la descripción, el resumen y el aprendizaje a partir de los datos. Resulta más adecuado para conjuntos de datos de tamaño moderado. (Si el tamaño del conjunto de datos fuera muy grande, desde un punto de vista práctico, los valores de las hojas podrían ser excesivos y puede que los gráficos de tallos y hojas no fueran más informativos que un histograma.) En cuanto a su forma, este gráfico se parece a un histograma girado, con el plus adicional de que presenta todos los valores existentes dentro de cada clase. Estos valores dentro de cada clase pueden ser de gran utilidad para detectar patrones en los datos, (tales como ver que todos los datos son múltiplos de algún valor), o para encontrar qué valores suceden con mayor frecuencia dentro de cada tallo.

En ocasiones, si un gráfico de tallos y hojas tiene demasiadas hojas por tallo resulta excesivamente desordenado. Una posible solución es la de duplicar el número de tallos, generando dos tallos nuevos por cada uno de los antiguos. Para los tallos del gráfico anterior, los pares de tallos nuevos podrían incluir todas las hojas con valores entre 0 y 4, por un lado, y los valores entre 5 y 9, por otro. Por ejemplo, supongamos que un tallo del gráfico antiguo fuera:

6 | 0, 0, 1, 2, 2, 3, 4, 4, 4, 4, 5, 5, 6, 6, 7, 7, 7, 7, 8, 9, 9

Éste se podría partir en los dos tallos siguientes:

6	0, 0, 1, 2, 2, 3, 4, 4, 4, 4
6	5, 5, 6, 6, 7, 7, 7, 7, 8, 9, 9

Problemas

1. Para los siguientes datos, dibuje los gráficos de tallos y hojas teniendo (a) 4 tallos y (b) 8 tallos.

   124, 129, 118, 135, 114, 139, 127, 141, 111, 144, 133, 127,
   122, 119, 132, 137, 146, 122, 119, 115, 125, 132, 118, 126,
   134, 147, 122, 119, 116, 125, 128, 130, 127, 135, 122, 141

2. Los datos siguientes muestran qué porcentajes de personas, con edad mayor o igual a 25 años, eran graduados universitarios, en el año 2002, para los distintos Estados de Estados Unidos y para el Distrito de Columbia. Represente estos datos mediante un gráfico de tallos y hojas.

Porcentajes estatales de personas de 25 años o más que son titulados universitarios, año 2002

Estado	Porcentaje	Ordinal*
Estados Unidos	**26,7**	**(X)**
Alabama	22,7	38
Alaska	25,6	26
Arizona	26,3	22
Arkansas	18,3	49
California	27,9	15
Colorado	35,7	2
Connecticut	32,6	5
Delaware	29,5	11
District of Columbia	44,4	(X)
Florida	25,7	25
Georgia	25,0	29
Hawaii	26,8	19
Idaho	20,9	45
Illinois	27,3	16
Indiana	23,7	33
Iowa	23,1	37
Kansas	29,1	12
Kentucky	21,6	43
Louisiana	22,1	41
Maine	23,8	32
Maryland	37,6	1
Massachusetts	34,3	4
Michigan	22,5	39
Minnesota	30,5	8
Mississippi	20,9	45
Missouri	26,7	21
Montana	23,6	34
North Dakota	25,3	28

Porcentajes estatales de personas de 25 años o más que son titulados universitarios, año 2002 (*Continuación*)

Estado	Porcentaje	Ordinal*
Ohio	24,5	31
Oklahoma	20,4	47
Oregon	27,1	17
Pennsylvania	26,1	24
Rhode Island	30,1	9
South Carolina	23,3	36
South Dakota	23,6	34
Tennessee	21,5	44
Texas	26,2	23
Utah	26,8	19
Vermont	30,8	7
Virginia	34,6	3
Washington	28,3	14
West Virginia	15,9	50
Wisconsin	24,7	30
Wyoming	19,6	48

* Cuando varios Estados comparten el mismo ordinal, se omite el siguiente valor ordinal. Puesto que se incluyen datos redondeados, los Estados pueden mostrar valores idénticos, siendo ligeramente distintos.
Fuente: Resumen Estadístico de Estados Unidos.

3. Los siguientes datos representan las edades, redondeadas al año más próximo, de 43 pacientes de emergencia en un hospital de adultos:

   23, 18, 31, 79, 44, 51, 24, 19, 17, 25, 27, 19, 44, 61, 22, 18,
   14, 17, 29, 31, 22, 17, 15, 40, 55, 16, 17, 19, 20, 32, 20, 45,
   53, 27, 16, 19, 22, 20, 18, 30, 20, 33, 21

   Construya un gráfico de tallos y hojas para este conjunto de datos. Utilice este gráfico para determinar el intervalo de edad de 5 años de amplitud que contiene el mayor número de datos.

4. Un psicólogo registró los 48 tiempos de reacción (en segundos) siguientes a un cierto estímulo.

   1,1, 2,1, 0,4, 3,3, 1,5, 1,3, 3,2, 2,0, 1,7, 0,6, 0,9, 1,6, 2,2, 2,6, 1,8, 0,9,
   2,5, 3,0, 0,7, 1,3, 1,8, 2,9, 2,6, 1,8, 3,1, 2,6, 1,5, 1,2, 2,5, 2,8, 0,7, 2,3,
   0,6, 1,8, 1,1, 2,9, 3,2, 2,8, 1,2, 2,4, 0,5, 0,7, 2,4, 1,6, 1,3, 2,8, 2,1, 1,5

   (a) Construya un gráfico de tallos y hojas para estos datos.

   (b) Construya un segundo gráfico de tallos y hojas usando tallos adicionales.

   (c) ¿Cuál de ellos parece más informativo?

   (d) Supóngamos que un artículo del periódico mantiene que "El tiempo típico de reacción es de_____ segundos". Rellene el hueco que falta con el número que debería figurar.

5. Los siguientes valores representan los ingresos diarios de los parquímetros de la ciudad de Nueva York (en unidades de 5000 dólares) en 30 días del año 2002.

   108, 77, 58, 88, 65, 52, 104, 75, 80, 83, 74, 68, 94, 97, 83,
   71, 78, 83, 90, 79, 84, 81, 68, 57, 59, 32, 75, 93, 100, 88

   (a) Represente estos datos mediante un gráfico de tallos y hojas.

   (b) ¿Parece algún dato "sospechoso"? ¿Por qué?

6. La volatilidad de una acción es una propiedad importante en la teoría de precios futuros. Representa un indicativo de la magnitud del cambio que, día a día, tiende a existir en los precios de la acción. Una volatilidad de 0 significa que el precio de la acción se mantiene constante. Cuanto más alta es la volatilidad, mayor es la tendencia al cambio del precio de las acciones. La lista siguiente muestra las volatilidades de 32 compañías cuyas acciones están incluidas en el mercado de valores de Estados Unidos:

   0,26, 0,31, 0,45, 0,30, 0,26, 0,17, 0,33, 0,32, 0,37, 0,38, 0,35, 0,28, 0,37,
   0,35, 0,29, 0,20, 0,33, 0,19, 0,31, 0,26, 0,24, 0,50, 0,22, 0,33, 0,51,
   0,44, 0,63, 0,30, 0,28, 0,48, 0,42, 0,37

   (a) Represente estos datos mediante un gráfico de tallos y hojas.

   (b) ¿Cuál es el dato de mayor magnitud?

   (c) ¿Cuál es el dato de menor valor?

   (d) ¿Cuál es el valor de dato "típico"?

7. La tabla siguiente muestra los resultados de los 25 primeros partidos de la Super Copa de fútbol profesional. Utilícela para construir un gráfico de tallos y hojas para

   (a) las puntuaciones ganadoras,

   (b) las puntuaciones perdedoras,

   (c) las cantidades que los puntos de los equipos ganadores sobrepasan a los de los equipos perdedores.

Super Bowls I-XXV

Partido	Fecha	Ganador	Perdedor
XXV	Jan. 27, 1991	Giants (NFC) 20	Buffalo (AFC) 19
XXIV	Jan. 28, 1990	San Francisco (NFC) 55	Denver (AFC) 10
XXIII	Jan. 22, 1989	San Francisco (NFC) 20	Cincinnati (AFC) 16
XXII	Jan. 31, 1988	Washington (NFC) 42	Denver (AFC) 10
XXI	Jan. 25, 1987	Giants (NFC) 39	Denver (AFC) 20
XX	Jan. 26, 1986	Chicago (NFC) 46	New England (AFC) 10
XIX	Jan. 20, 1985	San Francisco (NFC) 38	Miami (AFC) 16
XVIII	Jan. 22, 1984	Los Angeles Raiders (AFC) 38	Washington (NFC) 9
XVII	Jan. 30, 1983	Washington (NFC) 27	Miami (AFC) 17
XVI	Jan. 24, 1982	San Francisco (NFC) 26	Cincinnati (AFC) 21
XV	Jan. 25, 1981	Oakland (AFC) 27	Philadelphia (NFC) 10
XIV	Jan. 20, 1980	Pittsburgh (AFC) 31	Los Angeles (NFC) 19

Super Bowls I-XXV (*Continuación*)

Partido	Fecha	Ganador	Perdedor
XIII	Jan. 21, 1979	Pittsburgh (AFC) 35	Dallas (NFC) 31
XII	Jan. 15, 1978	Dallas (NFC) 27	Denver (AFC) 10
XI	Jan. 9, 1977	Oakland (AFC) 32	Minnesota (NFC) 14
X	Jan. 18, 1976	Pittsburgh (AFC) 21	Dallas (NFC) 17
IX	Jan. 12, 1975	Pittsburgh (AFC) 16	Minnesota (NFC) 6
VIII	Jan. 13, 1974	Miami (AFC) 24	Minnesota (NFC) 7
VII	Jan. 14, 1973	Miami (AFC) 14	Washington (NFC) 7
VI	Jan. 16, 1972	Dallas (NFC) 24	Miami (AFC) 3
V	Jan. 17, 1971	Baltimore (AFC) 16	Dallas (NFC) 13
IV	Jan. 11, 1970	Kansas City (AFL) 23	Minnesota (NFL) 7
III	Jan. 12, 1969	New York (AFL) 16	Baltimore (NFL) 7
II	Jan. 14, 1968	Green Bay (NFL) 33	Oakland (AFL) 14
I	Jan. 15, 1967	Green Bay (NFL) 35	Kansas City (AFL) 10

8. Considere el siguiente gráfico de tallos y hojas y el siguiente histograma referidos a un mismo conjunto de datos.

2	1, 1, 4, 7
3	0, 0, 3, 3, 6, 9, 9, 9
4	2, 2, 5, 8, 8, 8
5	1, 1, 7, 7
6	3, 3, 3, 6
7	2, 2, 5, 5, 5, 8

2–3	x, x, x, x
3–4	x, x, x, x, x, x, x, x
4–5	x, x, x, x, x, x
5–6	x, x, x, x
6–7	x, x, x, x
7–8	x, x, x, x, x, x

A partir del gráfico de tallos y hojas, ¿qué se puede concluir que no sea visible desde el histograma?

9. Utilice los datos representados en el gráfico de tallos y hojas del problema 8 para contestar a las siguientes cuestiones:

   (a) ¿Cuántos datos se encuentran en los 40?

   (b) ¿Qué porcentaje de valores se encuentran por encima de 50?

   (c) ¿Qué porcentaje de valores tiene el dígito de las unidades igual a 1?

10. La tabla siguiente muestra las tasas del impuesto sobre la renta y las tasas de la Seguridad Social correspondientes a 2002 para un cierto grupo de países.

    (a) Represente los porcentajes pagados por el impuesto sobre la renta mediante un histograma.

    (b) Represente los porcentajes pagados a la Seguridad Social mediante un histograma.

    (c) Represente los porcentajes pagados a la Seguridad Social mediante un gráfico de tallos y hojas.

Cargas fiscales en los países seleccionados*

País	Impuesto sobre la renta (%)	Seguridad Social (%)	Pago total† (%)	País	Impuesto sobre la renta (%)	Seguridad Social (%)	Pago total† (%)	País	Impuesto sobre la renta (%)	Seguridad Social (%)	Pago total†† (%)
Denmark	33	11	43	France	13	13	27	New Zealand	20	0	20
Belgium	28	14	41	Canada	19	7	26	Slovak Republic	7	13	19
Germany	21	21	41	Australia	24	0	24	Spain	13	6	19
Finland	26	6	32	Czech Republic	11	13	24	Greece	1	16	17
Poland	6	25	31	United States	17	8	24	Portugal	6	11	17
Sweden	23	7	30	United Kingdom	16	8	23	Ireland	11	5	16
Turkey	15	15	30	Iceland	22	0	22	Japan	6	10	16
Netherlands	7	22	29	Luxembourg	8	14	22	Korea	2	7	9
Norway	21	8	29	Switzerland	10	12	22	Mexico	2	2	4
Austria	11	18	29								
Hungary	17	13	29								
Italy	19	9	28								

* No se incluyen los impuestos no indicados, tales como los impuestos sobre las ventas o sobre el valor añadido. Las tasas mostradas se aplican a las personas individuales con rentas medias.
† Es posible que los totales no coincidan con la suma debido a los redondeos.
Fuente: Organización para la Cooperación y el Desarrollo Económico, 2002.

11. Una forma útil de comparar dos conjuntos de datos consiste en colocar sus gráficos de tallos y hojas contiguamente. A continuación se representan las calificaciones obtenidas por los estudiantes de dos escuelas distintas en un examen estándar. En ambas escuelas, 24 estudiantes se presentaron al examen.

Escuela A		Escuela B
Hojas	Tallo	Hojas
0	5	3, 5, 7
8, 5	6	2, 5, 8, 9, 9
9, 7, 4, 2, 0	7	3, 6, 7, 8, 8, 9
9, 8, 8, 7, 7, 6, 5, 3	8	0, 2, 3, 5, 6, 6
8, 8, 6, 6, 5, 5, 3, 0	9	0, 1, 5
	10	0

(a) ¿Qué escuela obtuvo la mayor calificación?

(b) ¿Qué escuela obtuvo la menor calificación?

(c) ¿Qué escuela obtuvo los mejores resultados?

(d) Reúna los datos de las dos escuelas y represente las 48 calificaciones mediante un gráfico de tallos y hojas.

2.5 Conjuntos de datos apareados

En ocasiones, los conjuntos de datos consisten en pares de valores con algún tipo de relación entre ellos. Cada individuo del conjunto de datos presenta un valor x y un valor y. Por lo general, el par i-ésimo se denota mediante (x_i, y_i), $i = 1,\ldots, n$. Por ejemplo, en el conjunto de datos presentado en la tabla 2.12, x_i representa la puntuación obtenida en el test de coeficiente de inteligencia (IQ), e y_i representa el salario anual (redondeado en miles de dólares) del i-ésimo trabajador de una muestra de 30 trabajadores pertenecientes a una empresa. En este apartado, se mostrará cómo se pueden representar de manera efectiva conjuntos de datos con valores apareados.

Una posibilidad de representación de esos conjuntos de datos consiste en considerar separadamente cada uno de los datos apareados y en representar cada uno de ellos mediante histogramas o gráficos de tallos y hojas. Por ejemplo, las figuras 2.11 y 2.12 muestran los gráficos de tallos y hojas, respectivamente, de las puntuaciones del test IQ y de los salarios anuales correspondientes a los datos incluidos en la tabla 2.12.

Sin embargo, aunque las figuras 2.11 y 2.12 exponen amplia información sobre las puntuaciones del test IQ y sobre los salarios de los trabajadores, no nos dicen nada acerca de la relación existente entre ambas variables. Así por ejemplo, no son útiles por sí mismas para ayudar a discernir si las mayores puntuaciones en el test de inteligencia tienden a corresponderse con los salarios más elevados de la compañía. Para responder a cuestiones de este tipo, es preciso considerar simultáneamente los valores apareados de cada dato puntual.

Una forma útil de mostrar un conjunto de datos con valores apareados es la de representarlos mediante un gráfico cartesiano con dos ejes perpendiculares. En el eje x aparecerían los valores x de los datos, mientras que los valores y estarían en el eje y. Tales gráficos se denominan *diagramas de dispersión*. La figura 2.13 presenta un diagrama de dispersión para los datos de la tabla 2.12.

Tabla 2.12 Salarios frente a puntuaciones del test IQ

Trabajador i	Puntuación IQ x_i	Salario anual y_i (en miles de dólares)	Trabajador i	Puntuación IQ x_i	Salario anual y_i (en miles de dólares)
1	110	68	16	84	19
2	107	30	17	83	16
3	83	13	18	112	52
4	87	24	19	80	11
5	117	40	20	91	13
6	104	22	21	113	29
7	110	25	22	124	71
8	118	62	23	79	19
9	116	45	24	116	43
10	94	70	25	113	44
11	93	15	26	94	17
12	101	22	27	95	15
13	93	18	28	104	30
14	76	20	29	115	63
15	91	14	30	90	16

12	4	(1)
11	0, 0, 2, 3, 3, 5, 6, 6, 7, 8	(10)
10	1, 4, 4, 7	(4)
9	0, 1, 1, 3, 3, 4, 4, 5	(8)
8	0, 3, 3, 4, 7	(5)
7	6, 9	(2)

Figura 2.11 Gráfico de tallos y hojas para las puntuaciones del test IQ.

7	0, 1	(2)
6	2, 3, 8	(3)
5	2	(1)
4	0, 3, 4, 5	(4)
3	0, 0	(2)
2	0, 2, 2, 4, 5, 9	(6)
1	1, 3, 3, 4, 5, 5, 6, 6, 7, 8, 9, 9	(12)

Figura 2.12 Gráfico de tallos y hojas para los salarios anuales (en miles de dólares).

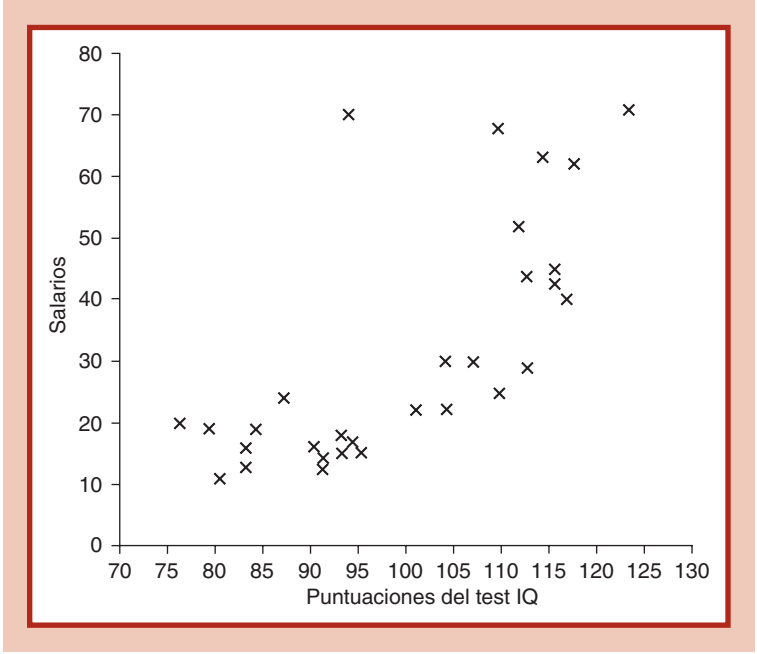


Figura 2.13 Diagrama de dispersión de puntuaciones del test IQ frente a los salarios.

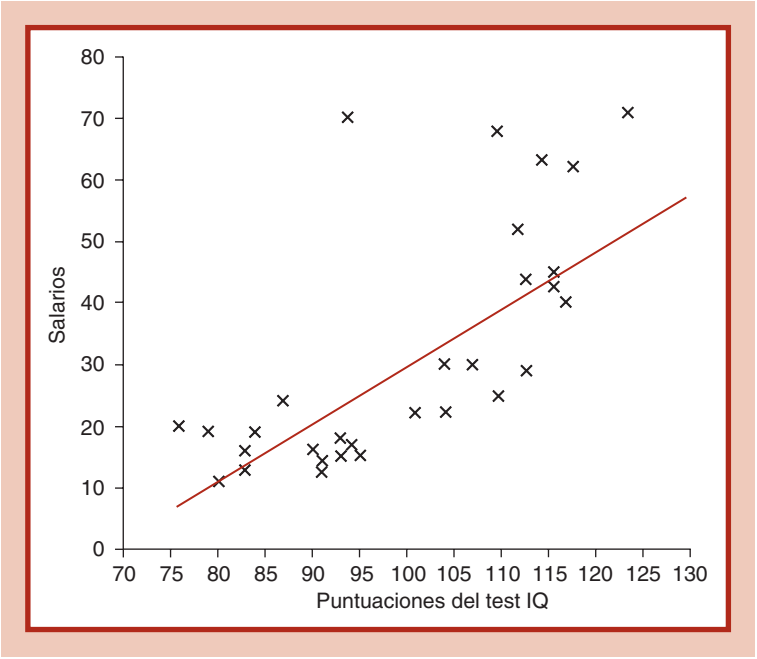


Figura 2.14 Diagrama de dispersión de puntuaciones del test IQ frente a los salarios; se ajusta a ojo una línea recta.

Resulta evidente, de la figura 2.13, que los salarios más altos se corresponden con las puntuaciones más altas del test IQ. Esto es, aunque no todos los trabajadores con puntuaciones IQ altas reciben un salario superior al que recibe otro trabajador con una puntuación menor (compare el trabajador 5 con el 29), generalmente, lo indicado resulta ser cierto.

El diagrama de dispersión de la figura 2.13 también puede resultar útil para establecer ciertos pronósticos. Por ejemplo, supongamos que pretendemos predecir el salario de un trabajador, similar a los considerados, cuya puntuación obtenida en el test de inteligencia fuera de 120. Una forma de hacerlo consiste en "ajustar a ojo" una línea recta al conjunto de datos, tal como se hizo en la figura 2.14. Puesto que el valor y en la recta correspondiente al valor x de 120 es más o menos de 45, este valor parece una predicción razonable para el salario de un trabajador cuya puntuación IQ sea 120.

Aparte de que representan los patrones conjuntos de dos variables y de que nos permiten hacer predicciones, los diagramas de dispersión resultan útiles para detectar *outliers*, los datos puntuales que aparentemente no siguen los patrones de los demás datos. [Por ejemplo, el punto (94,70) de la figura 2.13 parece que no siga la tendencia general.] Tras haber detectado los *outliers*, se puede decidir si el par de datos es significativo o si se debe a un error en la obtención de la información.

Problemas

1. Para determinar la relación entre la temperatura que hay al mediodía (medida en grados Celsius) y el número de piezas defectuosas producidas dicho día, una compañía registró los datos siguientes correspondientes a 22 días laborables.

Temperatura	Número de piezas defectuosas	Temperatura	Número de piezas defectuosas
24,2	25	24,8	23
22,7	31	20,6	20
30,5	36	25,1	25
28,6	33	21,4	25
25,5	19	23,7	23
32,0	24	23,9	27
28,6	27	25,2	30
26,5	25	27,4	33
25,3	16	28,3	32
26,0	14	28,8	35
24,4	22	26,6	24

(a) Dibuje un diagrama de dispersión.

(b) ¿Qué se puede concluir a partir del diagrama anterior?

(c) Si la temperatura al mediodía de mañana fuera de 24º C, ¿qué se podría conjeturar sobre el número de piezas defectuosas que se van a producir al día siguiente?

2. La tabla siguiente muestra, para cada Estado, el porcentaje de población que no dispone de seguro médico, en los años 1990, 2000 y 2002.

Cobertura* por seguros médicos en los Estados, 1990, 2000, 2002

	2002		2000		1990			2002		2000		1990	
	Sin seguro†	% sin seguro	Sin seguro†	% sin seguro	Sin seguro†	% sin seguro		Sin seguro†	% sin seguro	Sin seguro†	% sin seguro	Sin seguro†	% sin seguro
AL	564	12,7	582	13,3	710	17,4	MT	139	15,3	150	16,8	115	14,0
AK	119	18,7	117	18,7	77	15,4	NE	174	10,2	154	9,1	138	8,5
AZ	916	16,8	869	16,7	547	15,5	NV	418	19,7	344	16,8	201	16,5
AR	440	16,3	379	14,3	421	17,4	NH	125	9,9	103	8,4	107	9,9
CA	6 398	18,2	6 299	18,5	5 683	19,1	NJ	1 197	13,9	1 021	12,2	773	10,0
CO	720	16,1	620	14,3	495	14,7	NM	388	21,1	435	24,2	339	22,2
CT	356	10,5	330	9,8	226	6,9	NY	3 042	15,8	3 056	16,3	2 176	12,1
DE	79	9,9	72	9,3	96	13,9	NC	1 368	16,8	1 084	13,6	883	13,8
DC	74	13,0	78	14,0	109	19,2	ND	69	10,9	71	11,3	40	6,3
FL	2 843	17,3	2 829	17,7	2 376	18,0	OH	1 344	11,9	1 248	11,2	1 123	10,3
GA	1 354	16,1	1 166	14,3	971	15,3	OK	601	17,3	641	18,9	574	18,6
HI	123	10,0	113	9,4	81	7,3	OR	511	14,6	433	12,7	360	12,4
ID	233	17,9	199	15,4	159	15,2	PA	1 380	11,3	1 047	8,7	1 218	10,1
IL	1 767	14,1	1 704	13,9	1 272	10,9	RI	104	9,8	77	7,4	105	11,1
IN	797	13,1	674	11,2	587	10,7	SC	500	12,5	480	12,1	550	16,2

Cobertura* por seguros médicos en los Estados, 1990, 2000, 2002 (*Continuación*)

	2002		2000		1990			2002		2000		1990	
	Sin seguro†	% sin seguro	Sin seguro†	% sin seguro	Sin seguro†	% sin seguro		Sin seguro†	% sin seguro	Sin seguro†	% sin seguro	Sin seguro†	% sin seguro
IA	277	9,5	253	8,8	225	8,1	SD	85	11,5	81	11,0	81	11,6
KS	280	10,4	289	10,9	272	10,8	TN	614	10,8	615	10,9	673	13,7
KY	548	13,6	545	13,6	480	13,2	TX	5 556	25,8	4 748	22,9	3 569	21,1
LA	820	18,4	789	18,1	797	19,7	UT	310	13,4	281	12,5	156	9,0
ME	144	11,3	138	10,9	139	11,2	VT	66	10,7	52	8,6	54	9,5
MD	730	13,4	547	10,4	601	12,7	VA	962	13,5	814	11,6	996	15,7
MA	644	9,9	549	8,7	530	9,1	WA	850	14,2	792	13,5	557	11,4
MI	1 158	11,7	901	9,2	865	9,4	WV	255	14,6	250	14,1	249	13,8
MN	397	7,9	399	8,1	389	8,9	WI	538	9,8	406	7,6	321	6,7
MS	465	16,7	380	13,6	531	19,9	WY	86	17,7	76	15,7	58	12,5
MO	646	11,6	524	9,5	665	12,7	U,S,	43 574	15,2	39 804	14,2	34 719	13,9

* Para la población de todas las edades, incluidos los mayores de 65 años.
† En miles.
Fuente: Oficina de Censos, Estados Unidos. Departamento de Comercio.

(a) Dibuje un diagrama de dispersión en el que se relacionen las tasas correspondientes a los años 1990 y 2000.

(b) Dibuje un diagrama de dispersión para las tasas correspondientes a los años 2000 y 2002.

3. La tabla siguiente proporciona el número de habitantes de algunos de los condados más grandes de Estados Unidos.

Los 25 condados con mayor población, 2000–2002

Condado	2002 Población	2000 Población	Condado	2002 Población	2000 Población
Los Angeles, CA	9 806 577	9 519 330	Broward, FL	1 709 118	1 623 018
Cook, IL	5 377 507	5 376 741	Riverside, CA	1 699 112	1 545 387
Harris, TX	3 557 055	3 400 578	Santa Cara, CA	1 683 505	1 682 585
Maricopa, AZ	3 303 876	3 072 149	New York, NY	1 546 856	1 537 195
Orange, CA	2 938 507	2 846 289	Tarrant, TX	1 527 366	1 446 219
San Diego, CA	2 906 660	2 813 833	Clark, NV	1 522 164	1 375 738
Kings, NY	2 488 194	2 465 326	Philadelphia, PA	1 492 231	1 517 550
Miami-Dade, FL	2 332 599	2 253 362	Middlesex, MA	1 474 160	1 465 396
Dallas, TX	2 283 953	2 218 899	Alameda, CA	1 472 310	1 443 741
Queens, NY	2 237 815	2 229 379	Suffolk, NY	1 458 655	1 419 369
Wayne, MI	2 045 540	2 061 162	Bexar, TX	1 446 333	1 392 927
San Bemardino, CA	1 816 072	1 709 434	Cuyahoga, OH	1 379 049	1 393 845
King, WA	1 759 604	1 737 032			

Fuente: Oficina de Censos, Estados Unidos. Departamento de Comercio.

(a) Represente estos datos mediante un diagrama de dispersión.

(b) ¿Qué conclusiones se pueden sacar?

4. La tabla siguiente muestra el número de días en que, en los años comprendidos entre 1993 y 2002, no hubo los niveles de calidad aceptables en el aire de una muestra de distintas áreas metropolitanas de Estados Unidos.

Calidad del aire de las áreas metropolitanas de Estados Unidos seleccionadas, 1993–2002

Área metropolitana muestreada	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002
Atlanta, GA	36	15	36	28	33	52	67	34	18	24
Bakersfield, CA	97	105	107	110	58	78	144	132	125	152
Baltimore, MD	48	40	36	28	30	51	40	19	32	42
Boston, MA–NH	2	6	7	4	7	8	10	1	12	16
Chicago, IL	4	13	24	7	10	12	19	2	22	21
Dallas, TX	12	24	29	10	27	33	25	22	16	15
Denver, CO	6	3	5	2	0	9	5	3	8	8
Detroit, MI	5	11	14	13	11	17	20	15	27	26
El Paso, TX	7	6	3	6	2	6	5	4	9	13
Fresno, CA	59	55	61	70	75	67	133	131	138	152
Houston, TX	27	41	66	28	47	38	52	42	29	23
Las Vegas, NV–AZ	3	3	3	14	4	5	8	2	1	6
Los Angeles–Long Beach, CA	134	139	113	94	60	56	56	87	88	80
Miami, FL	6	1	2	1	3	8	7	2	1	1
Minneapolis–St. Paul, MN–WI	0	2	5	0	0	1	1	2	2	1
New Haven–Meriden, CT	12	13	14	8	19	9	19	9	15	25
New York, NY	11	16	21	14	23	18	25	19	19	31
Orange County, CA	25	15	9	9	3	6	14	31	31	19
Philadelphia, PA–NJ	62	37	38	38	38	37	32	22	29	33
Phoenix–Mesa, AZ	14	10	22	15	12	14	10	10	8	8
Pittsburgh, PA	14	22	27	12	21	39	40	29	52	53
Riverside–San Bernardino, CA	168	150	125	118	107	96	123	145	155	145
Sacramento, CA	20	37	41	44	17	29	69	45	49	69
St. Louis, MO–IL	9	33	38	23	15	24	31	18	17	34
Salt Lake City–Ogden, UT	5	17	5	14	2	19	8	15	15	18
San Diego, CA	59	46	48	31	14	33	33	31	31	20
San Francisco, CA	0	0	2	0	0	0	10	4	12	17
Seattle–Bellevue–Everett, WA	0	3	2	6	1	3	6	7	3	6
Ventura, CA	43	63	66	62	45	29	24	31	25	11
Washington, DC–MD–VA–WV	52	22	32	18	30	47	39	11	22	34

Nota: Los datos indican el número de días con niveles de calidad no aceptable en el aire de las áreas metropolitanas muestreadas. Todos los valores fueron revisados para ajustarse a los estándares de calidad establecidos en 1998. Son partículas aceptables las que tienen un diámetro menor o igual a 2,5 milímetros.
Fuente: Agencia de Protección Medioambiental de Estados Unidos, Oficina de Planificación y Estándares de Estados Unidos.

(a) Dibuje un diagrama de dispersión en el que se relacionen los valores de cada ciudad en los años 2000 y 2002.

(b) ¿Tienden a corresponderse los valores mayores del año 2002 con los valores mayores del año 2000?

5. Los datos siguientes relacionan el periodo de atención (en minutos) y la puntuación en un test de inteligencia (IQ) de 18 niños en edad preescolar.

Periodo de atención	Puntuación IQ	Periodo de atención	Puntuación IQ	Periodo de atención	Puntuación IQ
2,0	82	6,3	105	5,5	118
3,0	88	5,4	108	3,6	128
4,4	86	6,6	112	5,4	128
5,2	94	7,0	116	3,8	130
4,9	90	6,5	122	2,7	140
6,1	99	7,2	110	2,2	142

(a) Dibuje un diagrama de dispersión.

(b) Haga una inferencia plausible sobre la relación existente entre el periodo de atención y la puntuación IQ.

6. Los siguientes datos muestran los porcentajes de interés de los préstamos y las tasas de inflación durante 8 años de la década de 1970.

Tasa de inflación	Porcentaje de interés	Tasa de inflación	Porcentaje de interés
3,3	5,2	5,8	6,8
6,2	8,0	6,5	6,9
11,0	10,8	7,6	9,0
9,1	7,9		

(a) Dibuje un diagrama de dispersión.

(b) Ajuste a mano una recta para los pares de datos.

(c) A partir de la recta anterior, haga una predicción del porcentaje de interés de un año cuya tasa de inflación fuera del 7,2%.

7. Los datos siguientes muestran las rentas per cápita de los residentes de 12 áreas metropolitanas de Estados Unidos.

	Renta per cápita	
Área metropolitana	1994	1996
San Francisco–Oakland–San Jose, CA	28 990 $	32 933 $
Salt Lake City–Ogden, UT	18 731 $	21 271 $
Portland–Salem, OR–WA	22 508 $	25 343 $

Área metropolitana	Renta per cápita 1994	1996
Boston–Worcester–Lawrence–Lowell–Br ockton, MA–NH	27 095 $	30 366 $
Phoenix–Mesa, AZ	20 911 $	23 377 $
Seattle–Tacoma–Bremerton, WA	25 287 $	28 269 $
Denver–Boulder–Greeley, CO	25 657 $	28 650 $
Minneapolis–St. Paul, MN–WI	26 246 $	29 299 $
Tampa–St. Petersburg–Clearwater, FL	21 503 $	23 984 $
Charlotte–Gastonia–Rock Hill, NC–SC	22 819 $	25 446 $
Kansas City, MO–KS	23 281 $	25 949 $
Atlanta, GA	24 451 $	27 241 $

(a) Represente estos datos mediante un diagrama de dispersión.

(b) En 1994 la renta per cápita de los habitantes de San Diego, CA, fue de 22 111 dólares. Haga una predicción del valor correspondiente a 1996.

8. En el problema 7 de la sección 2.4 se muestran los resultados de los 25 primeros partidos de fútbol americano de la Super Copa. Para cada partido, y denota los puntos del equipo ganador y x denota el número de puntos en que este último equipo superó a su contrario. Dibuje un diagrama de dispersión en el que se relacionen x e y. ¿Tienden a corresponderse los valores mayores de una variable con los valores mayores de la otra?

2.6 Comentarios históricos

Probablemente el primer caso registrado de representaciones estadísticas –entiéndase, representaciones de datos mediante tablas y gráficos– se debe a Edmund Halley, con sus análisis gráficos de las presiones barométricas en función de la altitud. Se publicaron en 1686 y se utilizó el sistema de coordenadas cartesianas, introducido por el científico francés René Descartes en sus trabajos de geometría analítica. Halley presentó un diagrama de dispersión y también fue capaz de ajustar una curva a los puntos del gráfico.

A pesar del éxito que Halley consiguió con sus representaciones gráficas, hasta los últimos años del siglo XVIII la mayor parte de los científicos que trababan en esta materia prefirieron utilizar las tablas, en lugar de los gráficos, para presentar sus datos. En realidad, no fue hasta 1786, año en que William Playfair ideó el gráfico de barras como representación de una tabla de frecuencias, cuando se empezaron a utilizar regularmente las representaciones gráficas. En 1801, Playfair inventó los gráficos de tarta y, poco tiempo después, introdujo el uso de histogramas para visualizar datos.

El uso de gráficos para representar datos continuos –es decir, datos en los que todos los valores son distintos– no fue habitual hasta los años 1830. En 1833, el francés A. M. Guerry utilizó los gráficos de barras para representar datos sobre crímenes, tras haber clasificado los datos en intervalos para después reproducirlos en histogramas. En 1846, el estadístico y científico social Adolphe Quetelet hizo un uso sistemático de los histogramas. Quetelet y sus

estudiantes demostraron la utilidad del análisis gráfico en el desarrollo de las ciencias sociales. Tras ello, Quetelet popularizó la práctica, ampliamente extendida hoy día, de comenzar cualquier trabajo de investigación reuniendo primero los datos numéricos para representarlos después. Realmente, esta actuación, junto con los pasos adicionales de clasificación de los datos y de utilización de los métodos de la inferencia estadística para extraer conclusiones, se ha convertido en el paradigma aceptado para investigar en todas las áreas relacionadas con las ciencias sociales. Igualmente, se ha convertido en una técnica importante en otros campos, tales como la investigación médica (para contrastar nuevos medicamentos y terapias) y otras áreas tradicionalmente no numéricas como la literatura (para asignar autor) y la historia (tal como fue utilizada por el historiador francés F. Braudel).


(Princeton University Libraries)

John Tukey

El término histograma fue acuñado por Karlww Pearson en sus disertaciones sobre los gráficos estadísticos. En 1970, el estadístico de Estados Unidos John Tukey utilizó el diagrama de tallos y hojas, que puede interpretarse como una variante del histograma. En palabras de Tukey: "Mientras que el histograma utiliza una marca no cuantitativa para indicar un valor de datos, está claro que la mejor marca es un dígito."

Términos clave

Frecuencia: Número de veces en las que un valor dado aparece en un conjunto de datos.

Tabla de frecuencias: tabla que presenta, para un conjunto de datos dado, cada valor distinto junto con su frecuencia.

Gráfico de líneas: Gráfico de una tabla de frecuencias. La abscisa especifica un valor de dato, y la frecuencia de ocurrencia de tal valor se identifica con la altura de una línea horizontal.

Gráfico de barras (o diagrama de barras): Similar al gráfico de líneas, excepto en que la frecuencia de un valor coincide ahora con la altura de la barra.

Polígono de frecuencias: Gráfico de los valores distintos y sus frecuencias, en el que se conectan los puntos del gráfico mediante rectas.

Conjunto de datos simétrico: Un conjunto de datos es simétrico con respecto a un valor dado x_0 si las frecuencias de los valores $x_0 - c$ y x_0+c son iguales para todo valor de c.

Frecuencia relativa: Frecuencia de un valor dividida entre el número total de datos del conjunto de éstos.

Gráfico de tarta: Gráfico que representa las frecuencias relativas mediante la división de un círculo en sectores.

Histograma: Gráfico en el que los datos se dividen en intervalos de clase, cuyas frecuencias se muestran en un gráfico de barras.

Histograma de frecuencias relativas: Histograma en el que se muestran gráficamente las frecuencias relativas de cada dato del conjunto.

Gráfico de tallos y hojas: Similar a un histograma, con la excepción de que las frecuencias se indican en una lista con los últimos dígitos (las hojas) de los datos.

Diagrama de dispersión: Gráfico bidimensional de un conjunto de datos apareados.

Resumen

En este capítulo se han explicado distintas formas de representar gráficamente conjuntos de datos. Por ejemplo, consideremos el siguiente conjunto de 13 datos:

1, 2, 3, 1, 4, 2, 6, 2, 4, 3, 5, 4, 2

Se pueden representar estos valores mediante la siguiente tabla de frecuencias, que muestra cada valor distinto junto con el número de veces que aparece en el conjunto de datos:

Tabla de frecuencias

Valor	Frecuencia	Valor	Frecuencia
1	2	4	3
2	4	5	1
3	2	6	1

Los datos también se pueden visualizar gráficamente mediante un *gráfico de líneas*, o bien mediante un *gráfico de barras*. En ocasiones, los distintos valores de datos se representan mediante estos gráficos, y después los puntos resultantes se conectan mediante líneas rectas. Esto da lugar a un *polígono de frecuencias*.

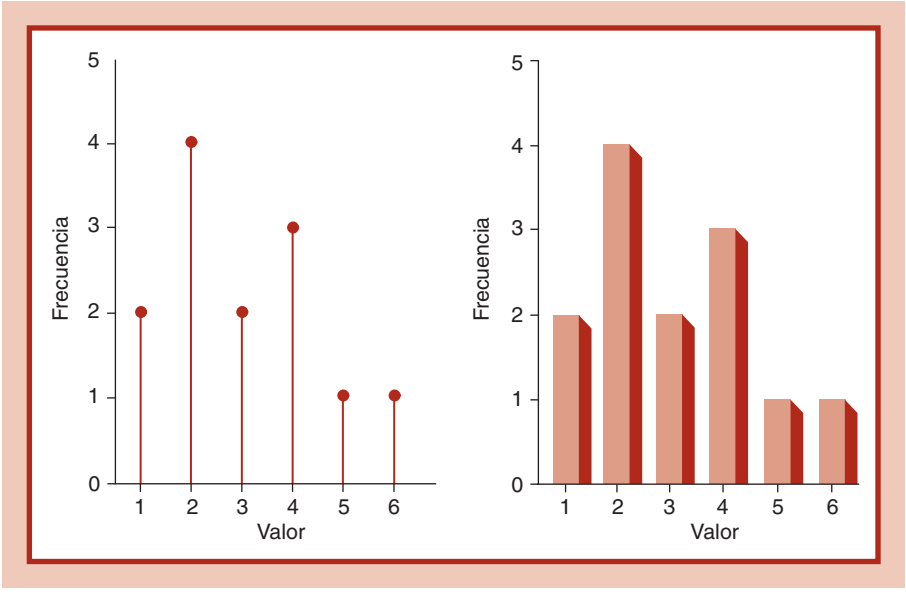
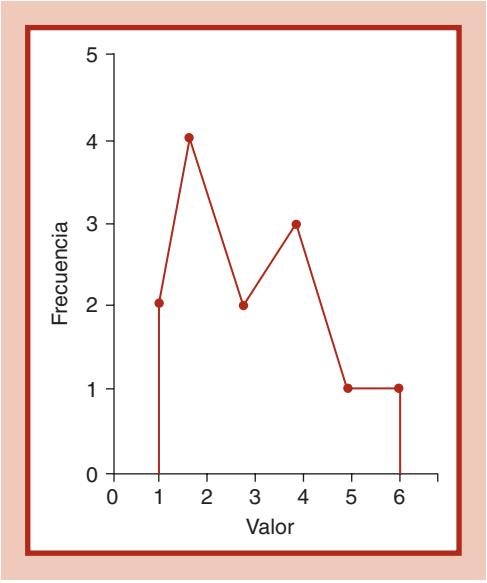
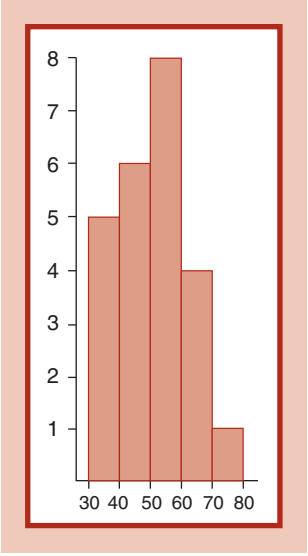


Gráfico de líneas. Gráfico de barras.



Polígono de frecuencias.

Cuando hay un gran número de valores de datos, éstos se suelen clasificar por intervalos de clase. Un gráfico de barras en el que se presentan los distintos intervalos de clase junto con el número de datos que aparecen dentro de cada intervalo se denomina *histograma*. En el eje *y* de este gráfico se pueden representar las frecuencias de clase (el número de valores dentro de cada intervalo de clase), o bien las proporciones de datos que aparecen dentro de cada clase. En el primer caso, el gráfico se denomina *histograma de frecuencias*; en el segundo, recibe el nombre de *histograma de frecuencias relativas*.



Histograma.

Considere el siguiente conjunto de datos:

41, 38, 44, 47, 33, 35, 55, 52, 41, 66, 64, 50, 49, 56,
55, 48, 52, 63, 59, 57, 75, 63, 38, 37

Si se usan los cinco intervalos de clase

30–40, 40–50, 50–60, 60–70, 70–80

junto con el convenio de inclusión por la izquierda (lo que significa que el intervalo contiene todos los puntos mayores o iguales que su extremo izquierdo y estrictamente menores que su extremo derecho), se consigue el histograma de la página 59 como representación del conjunto de datos citado.

Los conjuntos de datos también se pueden representar gráficamente mediante *gráficos de tallos y hojas*. El siguiente gráfico de tallos y hojas representa el anterior conjunto de datos.

7	5
6	3, 3, 4, 6
5	0, 2, 2, 5, 5, 6, 7, 9
4	1, 1, 4, 7, 8, 9
3	3, 5, 7, 8, 8

Gráfico de tallos y hojas.

En ocasiones, los datos se presentan en pares. Es decir, para cada elemento del conjunto de datos existe un valor x y un valor y. Un gráfico de los valores de x e y se denomina *diagrama de dispersión*. El diagrama de dispersión puede ser de gran utilidad para comprobar cuestiones tales como si los valores altos de x aparecen junto con valores altos de y, o si los valores altos de x se corresponden con valores bajos de y, o si no existe aparentemente ninguna relación entre los valores x e y de cada par.

El siguiente conjunto de pares de datos

i	1	2	3	4	5	6	7	8
x_i	8	12	7	15	5	12	10	22
y_i	14	10	17	9	13	8	12	6

se puede representar mediante el siguiente diagrama de dispersión. El diagrama indica que los valores altos de cualquier variable del par están, por lo general, asociados con los valores bajos de la otra variable.

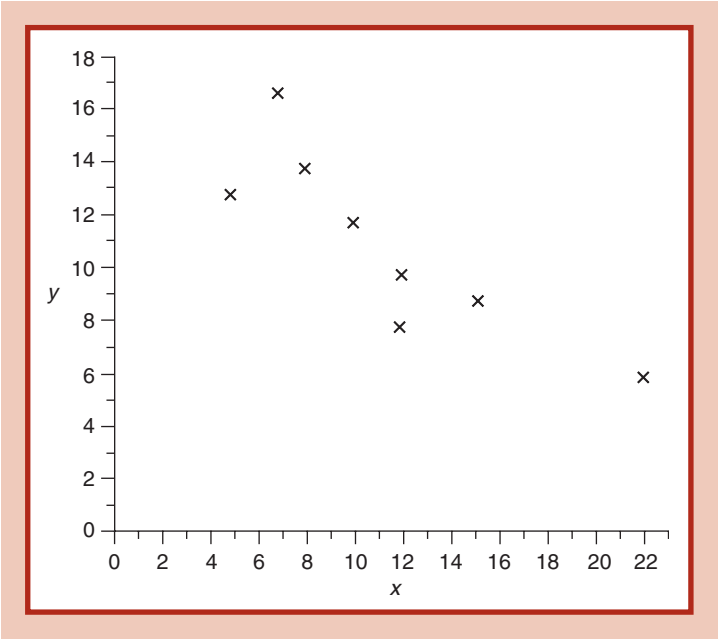


Diagrama de dispersión.

Normalmente, el uso de estas herramientas gráficas facilita que se reconozcan a primera vista las características relevantes de un conjunto de datos. Como resultado, se pueden poner de manifiesto aspectos que no resultan evidentes desde los propios datos en bruto. La elección de qué gráfico se va a utilizar depende de cuestiones tales como el tamaño del conjunto de datos, el tipo de datos o el número de valores distintos.

Problemas de repaso

1. Los siguientes datos muestran los tipos de sangre de 50 donantes voluntarios en cierta clínica:

   O A O AB A A O O B A O A AB B O O O A B A A O A A O
   B A O AB A O O A B A A A O B O O A O A B O AB A O B

   (a) Represente estos datos mediante una tabla de frecuencias.

   (b) Represéntelos también a través de una tabla de frecuencias relativas.

   (c) Represéntelos en un gráfico de tarta.

2. Los datos siguientes provienen de una muestra de precios, redondeados al céntimo más próximo, de un galón de gasolina estándar en el área de la Bahía de San Francisco en mayo de 1991.

   121, 119, 117, 121, 120, 120, 118, 124, 123, 139, 120,
   115, 117, 121, 123, 120, 123, 118, 117, 122, 122, 119

(a) Construya un histograma de frecuencias para este conjunto de datos.

(b) Construya un polígono de frecuencias.

(c) Construya un gráfico de tallos y hojas.

(d) ¿Existe algún dato aparentemente separado de los demás?

3. La siguiente tabla de frecuencias muestra el número de suicidios de mujeres, en ocho Estados alemanes durante 14 años.

Número de suicidios por año	0	1	2	3	4	5	6	7	8	9	10
Frecuencias	9	19	17	20	15	11	8	2	3	5	3

Así, por ejemplo, existieron 20 observaciones en las que ocurrieron 3 suicidios en los Estados y los años los correspondientes.

(a) ¿Cuántos suicidios se registraron a lo largo de los 14 años?

(b) Represente los datos anteriores mediante un histograma.

4. La tabla siguiente muestra las tasas de criminalidad (por 100 000 habitantes) de 1991 en los distintos Estados de Estados Unidos. Utilícela para construir:

(a) Un histograma de frecuencias de las tasas totales por crímenes violentos en los Estados nororientales.

(b) Un histograma de frecuencias relativas de las tasas totales por crímenes de propiedad en los Estados del sur.

(c) Un gráfico de tallos y hojas de las tasas por asesinato en los Estados del occidente.

(d) Un gráfico de tallos y hojas de las tasas por hurto en los Estados del oeste central.

		Crímenes violentos					Crímenes de propiedad			
Región, División, y Estado	**Total**	**Total**	**Asesinato**	**Secuestro con violencia**	**Robo**	**Atraco grave**	**Total**	**Robo**	**Hurto**	**Robo de automóvil**
Estados Unidos	**5 898**	**758**	**9,8**	**42**	**273**	**433**	**5 140**	**1 252**	**3 229**	**659**
Nororientales	5 155	752	8,4	29	352	363	4 403	1 010	2 598	795
New England	4 950	532	4,1	30	159	338	4 419	1 103	2 600	716
Maine	3 768	132	1,2	22	23	86	3 636	903	2 570	163
New Hampshire	3 448	119	3,6	30	33	53	3 329	735	2 373	220
Vermont	3 955	117	2,1	31	12	72	3 838	1 020	2 674	144
Massachusetts	5 332	736	4,2	32	195	505	4 586	1 167	2 501	919
Rhode Island	5 039	462	3,7	31	123	304	4 577	1 127	2 656	794
Connecticut	5 364	540	5,7	29	224	280	4 824	1 191	2 838	796
Atlántico Medio	5 227	829	9,9	29	419	372	4 398	978	2 598	823
New York	6 245	1 164	14,2	28	622	499	5 081	1 132	2 944	1 004
New Jersey	5 431	635	5,2	29	293	307	4 797	1 016	2 855	926
Pennsylvania	3 559	450	6,3	29	194	221	3 109	720	1 907	482

(Continuación)

Región, División, y Estado	Total	Crímenes violentos					Crímenes de propiedad			
		Total	Asesinato	Secuestro con violencia	Robo	Atraco grave	Total	Robo	Hurto	Robo de automóvil
Occidente central	5 257	631	7,8	45	223	355	4 626	1 037	3 082	507
Noreste central	5 482	704	8,9	50	263	383	4 777	1 056	3 151	570
Ohio	5 033	562	7,2	53	215	287	4 471	1 055	2 916	500
Indiana	4 818	505	7,5	41	116	340	4 312	977	2 871	465
Illinois	6 132	1 039	11,3	40	456	532	5 093	1 120	3 318	655
Michigan	6 138	803	10,8	79	243	470	5 335	1 186	3 469	680
Wisconsin	4 466	277	4,8	25	119	128	4 189	752	3 001	436
Noroeste central	4 722	457	5,4	34	129	288	4 265	991	2 918	356
Minnesota	4 496	316	3,0	40	98	175	4 180	854	2 963	363
Iowa	4 134	303	2,0	21	45	235	3 831	832	2 828	171
Missouri	5 416	763	10,5	34	251	467	4 653	1 253	2 841	558
North Dakota	2 794	65	1,1	18	8	38	2 729	373	2 229	127
South Dakota	3 079	182	1,7	40	19	122	2 897	590	2 192	115
Nebraska	4 354	335	3,3	28	54	249	4 020	727	3 080	213
Kansas	5 534	500	6,1	45	138	310	5 035	1 307	3 377	351
Sur	6 417	798	12,1	45	252	489	5 618	1 498	3 518	603
Atlántico sur	6 585	851	11,4	44	286	510	5 734	1 508	3 665	561
Delaware	5 869	714	5,4	86	215	408	5 155	1 128	3 652	375
Maryland	6 209	956	11,7	46	407	492	5 253	1 158	3 365	731
District of Columbia	10 768	2 453	80,6	36	1 216	1 121	8 315	2 074	4 880	1 360
Virginia	4 607	373	9,3	30	138	196	4 234	783	3 113	339
West Virginia	2 663	191	6,2	23	43	119	2 472	667	1 631	175
North Carolina	5 889	658	11,4	35	178	434	5 230	1 692	3 239	299
South Carolina	6 179	973	11,3	59	171	731	5 207	1 455	3 365	387
Georgia	6 493	738	12,8	42	268	415	5 755	1 515	3 629	611
Florida	8 547	1 184	9,4	52	400	723	7 363	2 006	4 573	784
Sureste central	4 687	631	10,4	41	149	430	4 056	1 196	2 465	395
Kentucky	3 358	438	6,8	35	83	313	2 920	797	1 909	215
Tennessee	5 367	726	11,0	46	213	456	4 641	1 365	2 662	614
Alabama	5 366	844	11,5	36	153	644	4 521	1 269	2 889	363
Mississippi	4 221	389	12,8	46	116	214	3 832	1 332	2 213	286
Suroeste central	7 118	806	14,2	50	254	488	6 312	1 653	3 871	788
Arkansas	5 175	593	11,1	45	136	402	4 582	1 227	3 014	341
Louisiana	6 425	951	16,9	41	279	614	5 473	1 412	3 489	573
Oklahoma	5 669	584	7,2	51	129	397	5 085	1 478	3 050	557
Texas	7 819	840	15,3	53	286	485	6 979	1 802	4 232	944
Oeste	6 478	841	9,6	46	287	498	5 637	1 324	3 522	791
Mountain	6 125	544	6,5	44	122	371	5 581	1 247	3 843	491
Montana	3 648	140	2,6	20	19	99	3 508	524	2 778	206
Idaho	4 196	290	1,8	29	21	239	3 905	826	2 901	178

(Continuación)

Región, División, y Estado	Total	Crímenes violentos					Crímenes de propiedad			
		Total	Asesinato	Secuestro con violencia	Robo	Atraco grave	Total	Robo	Hurto	Robo de automóvil
Wyoming	4 389	310	3,3	26	17	264	4 079	692	3 232	155
Colorado	6 074	559	5,9	47	107	399	5 515	1 158	3 930	426
New Mexico	6 679	835	10,5	52	120	652	5 845	1 723	3 775	346
Arizona	7 406	671	7,8	42	166	455	6 735	1 607	4 266	861
Utah	5 608	287	2,9	46	55	183	5 321	840	4 240	241
Nevada	6 299	677	11,8	66	312	287	5 622	1 404	3 565	652
Pacífico	6 602	945	10,7	47	345	542	5 656	1 351	3 409	896
Washington	6 304	523	4,2	70	146	303	5 781	1 235	4 102	444
Oregon	5 755	506	4,6	53	150	298	5 249	1 176	3 598	474
California	6 773	1 090	12,7	42	411	624	5 683	1 398	3 246	1 039
Alaska	5 702	614	7,4	92	113	402	5 088	979	3 575	534
Hawaii	5 970	242	4,0	33	87	118	5 729	1 234	4 158	336

Fuente: Oficina Federal de Investigación de Estados Unidos, Crimen en Estados Unidos, anuario.

5. Construya una tabla de frecuencias para un conjunto de datos de 10 valores que sea simétrico y tenga (a) 5 valores distintos y (b) 4 valores distintos. (c) ¿Con respecto a qué valores son simétricos los conjuntos de datos de los apartados (a) y (b)?

6. Los datos siguientes se refieren a las reservas de petróleo estimadas, en miles de millones de barriles, en cuatro regiones del hemisferio occidental. Represente estos datos mediante un gráfico de tarta.

Estados Unidos	38,7
América del Sur	22,6
Canadá	8,8
México	60,0

7. La siguiente tabla contiene las cantidades (en millones de dólares) invertidas en Estados Unidos procedentes de una selección de países europeos en los años 2000 y 2002.

Inversión extranjera en Estados Unidos, procedente de una selección de países (en millones de dólares)

	2000	2002
Europa	887 014	1 006 530
Austria	3 007	3 439
Bélgica	14 787	9 608
Dinamarca	4 025	1 924
Finlandia	8 875	7 212
Francia	125 740	170 619

Inversión extranjera en Estados Unidos, procedente de una selección de países (en millones de dólares)

	2000	2002
Alemania	122 412	137 036
Irlanda	25 523	26 179
Italia	6 576	6 695
Liechtenstein	319	259
Luxemburgo	58 930	34 349
Holanda	138 894	154 753
Noruega	2 665	3 416
España	5 068	4 739
Suecia	21 991	21 989
Suiza	64 719	113 232
Reino Unido	277 613	283 317

Fuente: Oficina de Análisis Económico, Departamento de Comercio de Estados Unidos.

(a) Represente los datos de 2000 y 2002 en dos gráficos de tarta contiguos.

(b) Dibuje su diagrama de dispersión.

8. Los datos siguientes se refieren a las edades (redondeadas al entero más próximo) en las que fallecieron cierto número de pacientes de un hospital (sin servicio de natalidad) de una gran ciudad:

1, 1, 1, 1, 3, 3, 4, 9, 17, 18, 19, 20, 20, 22, 24, 26, 28, 34,
45, 52, 56, 59, 63, 66, 68, 68, 69, 70, 74, 77, 81, 90

(a) Represente este conjunto de datos en un histograma.

(b) Represéntelo mediante un polígono de frecuencias.

(c) Represéntelo mediante un polígono de frecuencias relativas.

(d) Represéntelo en un gráfico de tallos y hojas.

Los problemas del 9 al 11 se refieren a los últimos 50 estudiantes del Apéndice A.

9. (a) Dibuje un histograma de los pesos de estos estudiantes.

(b) Comente ese histograma.

10. Dibuje un diagrama de dispersión que relacione los pesos con los niveles de colesterol. Comente qué se refleja en ese diagrama.

11. Dibuje un diagrama de dispersión que relacione los pesos y las presiones sanguíneas. ¿Qué le sugiere ese diagrama?

Los problemas 12 y 13 se refieren a la tabla siguiente, donde se muestran las calificaciones en Matemáticas y Lengua de varios estudiantes del último curso de educación secundaria.

Estudiante	Calificación en Lengua	Calificación en Matemáticas	Estudiante	Calificación en Lengua	Calificación en Matemáticas
1	520	505	8	620	576
2	605	575	9	604	622
3	528	672	10	720	704
4	720	780	11	490	458
5	630	606	12	524	552
6	504	488	13	646	665
7	530	475	14	690	550

12. Dibuje dos gráficos de tallos y hojas contiguos para las calificaciones de Matemáticas y Lengua. ¿Los estudiantes, como grupo, han obtenido las mejores calificaciones en una de estas asignaturas? Si es así, ¿en cuál?

13. Dibuje un diagrama de dispersión para las calificaciones de los estudiantes en ambas materias. ¿Tienden a aparecer las calificaciones altas en una asignatura junto a las calificaciones altas en la otra?

14. La tabla siguiente proporciona información acerca de las edades de los habitantes de Estados Unidos y México.

	Proporción de población (en porcentaje)	
Edad	México	Estados Unidos
0–9	32,5	17,5
10–19	24	20
20–29	14,5	14,5
30–39	11	12
40–49	7,5	12,5
50–59	4,5	10,5
60–69	3,5	7
70–79	1,5	4
Más de 80	1	2

(a) ¿Qué porcentaje de la población de México tiene menos de 30 años?

(b) ¿Qué porcentaje de la población de Estados Unidos tiene menos de 30 años?

(c) Dibuje dos polígonos de frecuencias relativas en un mismo gráfico. Utilice colores distintos para los datos de México y de Estados Unidos.

(d) En general, ¿cómo compararía las distribuciones de edad de ambos países?

15. Los datos siguientes se refieren a las precipitaciones anuales y mensuales (en pulgadas) que son habituales en varias ciudades.

Precipitaciones habituales, mensuales y anuales, en las ciudades seleccionadas

Estado	Ciudad	Ene.	Feb.	Mar.	Abr.	May.	Jun.	Jul.	Ago.	Sep.	Oct.	Nov.	Dic.	Anual
AL	Mobile	4,59	4,91	6,48	5,35	5,46	5,07	7,74	6,75	6,56	2,62	3,67	5,44	64,64
AK	Juneau	3,69	3,74	3,34	2,92	3,41	2,98	4,13	5,02	6,40	7,71	5,15	4,66	53,15
AZ	Phoenix	0,73	0,59	0,81	0,27	0,14	0,17	0,74	1,02	0,64	0,63	0,54	0,83	7,11
AR	Little Rock	3,91	3,83	4,69	5,41	5,29	3,67	3,63	3,07	4,26	2,84	4,37	4,23	49,20
CA	Los Angeles	3,06	2,49	1,76	0,93	0,14	0,04	0,01	0,10	0,15	0,26	1,52	1,62	12,08
	Sacramento	4,03	2,88	2,06	1,31	0,33	0,11	0,05	0,07	0,27	0,86	2,23	2,90	17,10
	San Diego	2,11	1,43	1,60	0,78	0,24	0,06	0,01	0,11	0,19	0,33	1,10	1,36	9,32
	San Francisco	4,65	3,23	2,64	1,53	0,32	0,11	0,03	0,05	0,19	1,06	2,35	3,55	19,71
CO	Denver	0,51	0,69	1,21	1,81	2,47	1,58	1,93	1,53	1,23	0,98	0,82	0,55	15,31
CT	Hartford	3,53	3,19	4,15	4,02	3,37	3,38	3,09	4,00	3,94	3,51	4,05	4,16	44,39
DE	Wilmington	3,11	2,99	3,87	3,39	3,23	3,51	3,90	4,03	3,59	2,89	3,33	3,54	41,38
DC	Washington	2,76	2,62	3,46	2,93	3,48	3,35	3,88	4,40	3,22	2,90	2,82	3,18	39,00
FL	Jacksonville	3,07	3,48	3,72	3,32	4,91	5,37	6,54	7,15	7,26	3,41	1,94	2,59	52,76
	Miami	2,08	2,05	1,89	3,07	6,53	9,15	5,98	7,02	8,07	7,14	2,71	1,86	57,55
GA	Atlanta	4,91	4,43	5,91	4,43	4,02	3,41	4,73	3,41	3,17	2,53	3,43	4,23	48,61
HI	Honolulu	3,79	2,72	3,48	1,49	1,21	0,49	0,54	0,60	0,62	1,88	3,22	3,43	23,47
ID	Boise	1,64	1,07	1,03	1,19	1,21	0,95	0,26	0,40	0,58	0,75	1,29	1,34	11,71
IL	Chicago	1,60	1,31	2,59	3,66	3,15	4,08	3,63	3,53	3,35	2,28	2,06	2,10	33,34
	Peoria	1,60	1,41	2,86	3,81	3,84	3,88	3,99	3,39	3,63	2,51	1,96	2,01	34,89
IN	Indianapolis	2,65	2,46	3,61	3,68	3,66	3,99	4,32	3,46	2,74	2,51	3,04	3,00	39,12
IA	Des Moines	1,01	1,12	2,20	3,21	3,96	4,18	3,22	4,11	3,09	2,16	1,52	1,05	30,83
KS	Wichita	0,68	0,85	2,01	2,30	3,91	4,06	3,62	2,80	3,45	2,47	1,47	0,99	28,61
KY	Louisville	3,38	3,23	4,73	4,11	4,15	3,60	4,10	3,31	3,35	2,63	3,49	3,48	43,56
LA	New Orleans	4,97	5,23	4,73	4,50	5,07	4,63	6,73	6,02	5,87	2,66	4,06	5,27	59,74

Fuente: Administración Nacional Oceánica y Atmosférica de Estados Unidos, *Climatología de Estados Unidos*, Septiembre, 1982.

(a) Represente las precipitaciones habituales del mes de abril en un gráfico de tallos y hojas.

(b) Represente las cantidades anuales en un histograma.

(c) Dibuje un diagrama de dispersión que relacione las cantidades de abril con las anuales.

16. Un valor muy separado del resto de valores se llama *outlier* (o valor extremo). En los siguientes conjuntos de datos, especifique qué valores son *outliers*, si es que existen.

(a) 14, 22, 17, 5, 18, 22, 10, −17, 25, 28, 33, 12

(b) 5, 2, 13, 16, 9, 12, 7, 10, 54, 22, 18, 15, 12

(c) 18, 52, 14, 20, 24, 27, 43, 17, 25, 28, 3, 22, 6

17. En la siguiente tabla se presentan datos sobre el número de coches importados en Estados Unidos procedentes de Japón y Alemania, en los años comprendidos entre 1970 y 2002.

Coches nuevos importados en Estados Unidos

	Japón	Alemania
1970	381 338	674 945
1971	703 672	770 807
1972	697 788	676 967
1973	624 805	677 465
1974	791 791	619 757
1975	695 573	370 012
1976	1 128 936	349 804
1977	1 341 530	423 492
1978	1 563 047	416 231
1979	1 617 328	495 565
1980	1 991 502	338 711
1981	1 911 525	234 052
1982	1 801 185	259 385
1983	1 871 192	239 807
1984	1 948 714	335 032
1985	2 527 467	473 110
1986	2 618 711	451 699
1987	2 417 509	377 542
1988	2 123 051	264 249
1989	2 051 525	216 881
1990	1 867 794	245 286
1991	1 762 347	171 097
1992	1 598 919	205 248
1993	1 501 953	180 383
1994	1 488 159	178 774
1995	1 114 360	204 932
1996	1 190 896	234 909
1997	1 387 812	300 489
1998	1 456 081	373 330
1999	1 707 277	461 061
2000	1 839 093	488 323
2001	1 790 346	494 131
2002	2 046 902	574 455

Fuente: Oficina de Censos, División de Comercio Exterior.

(a) ¿Qué conclusiones se pueden sacar acerca del número de coches alemanes y japoneses importados en Estados Unidos desde 1990?

(b) Presente un diagrama de dispersión que relacione las importaciones de coches japoneses y alemanes desde 1990.

3

Uso de la Estadística para sintetizar conjuntos de datos

Odio los promedios. No se puede cometer mayor error que decir que la Aritmética es una ciencia exacta. Existen permutaciones y aberraciones discernibles para mentes perfectamente nobles como la mía; cambios sutiles que los contables normales no pueden descubrir, escondidas leyes de los números que sólo pueden ser percibidas por una mente como la mía. Por ejemplo, si se promedian números de abajo a arriba y después de arriba abajo, el resultado es siempre distinto.

Carta a la *Mathematical Gazette*
(revista matemática británica del siglo XIX)

La forma de dar sentido a los datos en bruto consiste en comparar y contrastar, para entender las diferencias.

Gregory Bateson, en *Pasos hacia una ecología de la mente*

El objetivo de este capítulo es desarrollar medidas que se puedan usar para sintentizar un conjunto de datos. Estas medidas, formalmente llamadas *estadísticos*, son magnitudes numéricas cuyos valores vienen determinados por los datos. Se estudiarán la media muestral, la mediana muestral y la moda muestral, estadísticos que miden el centro o el valor central de un conjunto de datos. También se considerarán otros estadísticos que informan sobre la variación del conjunto de datos. Se aprenderá qué sig-

nifica que un conjunto de datos sea normal, y se presentará una regla empírica relativa a los conjuntos normales. También se estudiarán los conjuntos de datos compuestos por valores apareados, y se presentará un estadístico que sirve para medir el grado en el que un diagrama de dispersión de datos apareados se puede aproximar por una recta.

3.1 Introducción

En los experimentos actuales a menudo se hace un seguimiento de miles de individuos, y se observan algunas de sus características a lo largo del tiempo. Por ejemplo, en 1951, para conocer qué consecuencias en la salud se derivan de ciertas prácticas habituales, los médicos estadísticos R. Doll y A. B. Hill enviaron unos cuestionarios a todos los médicos de Reino Unido, y recibieron las respuestas de 40 000 médicos. En los cuestionarios se solicitaba información sobre la edad, los hábitos alimentarios y deportivos y sobre el consumo de tabaco. A esos médicos se les hizo un seguimiento durante 10 años, y se determinó la causa de muerte de los que fallecieron durante ese periodo de control. Como se puede imaginar, en ese estudio se utilizó un extensísimo conjunto de datos. Por ejemplo, aunque la atención se centrara en una única variable, tal como la edad de los doctores, en un determinado periodo de tiempo, el conjunto de datos resultante tendría un número de valores muy elevado: 40 000. Para que se pueda intuir la información contenida en un conjunto de datos tan grande es necesario resumir o sintetizar el conjunto de datos mediante una serie de medidas. En este capítulo se mostrarán los distintos estadísticos que se pueden utilizar para sintetizar determinadas características de un conjunto de datos.

Para empezar, supongamos que se dispone de un conjunto de datos muestrales procedentes de una población subyacente. Mientras que en capítulo 2 se mostró cómo describir y representar gráficamente los conjuntos de datos en toda su extensión, aquí nos interesaremos en determinar ciertas medidas sintéticas referidas a los datos. Estas medidas reciben el nombre de *estadísticos*. Se entiende por *estadístico* cualquier magnitud numérica cuyo valor se pueda determinar a partir de los datos.

Definición

Cualquier magnitud numérica calculable a partir de los datos se denomina *estadístico*.

Nos fijaremos en los estadísticos que describen la tendencia central del conjunto de datos; es decir, que describen el centro del conjunto de valores de datos. En las secciones 3.2, 3.3 y 3.4 se presentarán sucesivamente tres estadísticos de este tipo: la media muestral, la mediana muestral y la moda muestral. Una vez que se tenga una idea sobre el centro de un conjunto de datos, la cuestión que surge de manera natural es qué cantidad de *variación* existe. Esto es, ¿la mayor parte de los valores están próximos al centro, o, por el contrario, varían mucho alrededor de éste? En la sección 3.5 se analizarán la varianza muestral y la desviación típica muestral, que son estadísticos diseñados para medir la variación.

En la sección 3.6 se introducirá el concepto de conjunto normal de datos, áquel cuyo histograma tiene una forma acampanada. Para los conjuntos de datos próximos a la normalidad, se presentará una regla que se puede utilizar para aproximar la proporción de datos que distan de la media muestral menos de un cierto número de veces la desviación típica.

En las seis primeras secciones de este capítulo se tratan conjuntos de datos en los que cada dato está compuesto por un solo valor. Por otra parte, en la sección 3.7 se tratarán los datos apareados. Esto es, cada dato puntual consistirá en un valor x y un valor y. Por ejemplo, el valor x podría representar el número medio de cigarrillos que un fumador consume al día, mientras que el valor y podría identificarse con la edad de fallecimiento del individuo. Se introducirá un estadístico denominado *coeficiente de correlación muestral*, cuyo valor indica el grado en el que los datos puntuales con valores altos de x presentan, igualmente, valores altos de y; o, alternativamente, el grado en que valores bajos de x van unidos a valores bajos de y.

Del estudio de Doll y Hill se deduce que sólo alrededor del 1 por 1000 de los doctores no fumadores falleció de cáncer de pulmón. Entre los fumadores compulsivos, la cifra fue de 1 de cada 8. Adicionalmente, la tasa de mortalidad por ataque de corazón para los fumadores resultó ser un 50% más alta que para los no fumadores.

3.2 Media muestral

Supongamos que se dispone de una muestra de n datos cuyos valores serán designados por $x_1, x_2, \ldots, x_n$. Un estadístico usado para indicar el centro de este conjunto de datos es la *media muestral*, definida como la media aritmética de los valores de datos.

Definición

La *media muestral*, denotada por $\bar{x}$ (léase, "x barra"), se define por:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Ejemplo 3.1 Las eficiencias en el consumo de carburante (medidas en número de millas recorridas por galón de carburante) de los coches vendidos en Estados Unidos durante los años comprendidos entre 1999 y 2003 tuvieron como promedio:

28,2, 28,3, 28,4, 28,5, 29,0

Encuentre la media de este conjunto de datos.

Solución La media muestral $\bar{x}$ coincide con la media aritmética de los cinco valores de datos. Así pues,

$$\bar{x} = \frac{28{,}2 + 28{,}3 + 28{,}4 + 28{,}5 + 29{,}0}{5} = \frac{142{,}4}{5} = 28{,}48$$

Observe a partir de este ejemplo que, aunque la media muestral es la media aritmética de los distintos valores de datos, no tiene por qué coincidir con ninguno de éstos. ■

Consideremos de nuevo el conjunto de datos $x_1, x_2, \ldots, x_n$. Si cada valor de dato se incrementa en una constante c, la media muestral se incrementa igualmente en el valor c. Matemáticamente, esto se puede expresar diciendo que si

$$y_i = x_i + c \quad \text{para} \quad i = 1, \ldots, n$$

se verifica que

$$\bar{y} = \bar{x} + c$$

donde $\bar{y}$ y $\bar{x}$ representan las medias muestrales de los valores y_i y de los valores x_i, respectivamente. Por consiguiente, cuando sea conveniente, se puede calcular $\bar{x}$ si se añade, primero, c a todos los valores de datos; después, se calcula la media muestral $\bar{y}$; y, finalmente, se resta c a $\bar{y}$ para obtener $\bar{x}$. Puesto que en ocasiones es más sencillo trabajar con los datos transformados en lugar de con los datos originales, el proceso indicado puede simplificar enormemente el cálculo de $\bar{x}$. El siguiente ejemplo ilustra este hecho.

Ejemplo 3.2 Las puntuaciones obtenidas por los ganadores del Torneo de Maestros de Golf de Estados Unidos entre 1981 y 1990 fueron las siguientes:

280, 284, 280, 277, 282, 279, 285, 281, 283, 278

Encuentre la media muestral de las puntuaciones anteriores.

Solución En vez de sumar directamente todos los valores anteriores, restemos primero 280 (esto es, $c = -280$) de cada uno de ellos. Así se obtienen los siguientes datos transformados:

$$0, 4, 0, -3, 2, -1, 5, 1, 3, -2$$

La media muestral $\bar{y}$, de estos últimos valores es

$$\bar{y} = \frac{0 + 4 + 0 - 3 + 2 - 1 + 5 + 1 + 3 - 2}{10} = \frac{9}{10}$$

Si a $\bar{y}$ le añadimos 280 se obtiene que la media de los datos originales es

$$\bar{x} = 280{,}9 \quad \blacksquare$$

Si cada valor de dato se multiplica por c, igualmente queda multiplicada por c la media resultante. Esto es, si

$$y_i = cx_i \quad i = 1, \ldots, n$$

se verifica que

$$\bar{y} = c\bar{x}$$

Por ejemplo, supongamos que la media de las alturas de un conjunto de individuos es de 5 pies. Si se quisiera cambiar la unidad de medida de pies a pulgadas, cada nuevo valor coin-

cidiría con el antiguo multiplicado por 12. Se sigue que la media muestral de los datos nuevos coincide con $12 \cdot 5 = 60$. Es decir, la media muestral es de 60 pulgadas.

En el siguiente ejemplo se aborda el cálculo de la media muestral cuando los datos vienen dados mediante una tabla de frecuencias.

Ejemplo 3.3 El número de vestidos vendidos diariamente en una boutique de señoras durante los seis últimos días viene expresado en la tabla de frecuencias siguiente:

Valor	Frecuencia
3	2
4	1
5	3

¿Cuál es la media muestral?

Solución Dado que el conjunto de datos originales se compone de los siguientes 6 valores

$$3, 3, 4, 5, 5, 5$$

la media muestral resultante será

$$\begin{aligned} \bar{x} &= \frac{3 + 3 + 4 + 5 + 5 + 5}{6} \\ &= \frac{3 \times 2 + 4 \times 1 + 5 \times 3}{6} \\ &= \frac{25}{6} \end{aligned}$$

Esto es, la media muestral del número de vestidos vendidos diariamente es de 4,25. ■

En el ejemplo 3.3 se ha visto que, cuando los datos se dan mediante una tabla de frecuencias, la media muestral se puede expresar como la suma de los productos de los valores distintos y sus frecuencias dividida por el tamaño del conjunto de datos. Este resultado se verifica siempre. Para verlo, supongamos que los datos vienen dados en una tabla de frecuencias, donde se incluyen los k valores distintos, $x_1, x_2, \ldots, x_k$, junto con sus respectivas frecuencias, $f_1, f_2, \ldots, f_k$. El conjunto de datos consistirá, pues, en n observaciones, donde $n = \sum_{i=1}^{k} f_i$ y donde cada valor x_i aparece f_i veces, para $i = 1, \ldots, k$. Por consiguiente, la media muestral de este conjunto de datos será

$$\begin{aligned} \bar{x} &= \frac{x_1 + \cdots + x_1 + x_2 + \cdots + x_2 + \cdots + x_k + \cdots + x_k}{n} \\ &= \frac{f_1 x_1 + f_2 x_2 + \cdots + f_k x_k}{n} \end{aligned} \tag{3.1}$$

Si $w_1, w_2, \ldots, w_k$ son números positivos que sumen 1, la suma

$$w_1x_1 + w_2x_2 + \cdots + w_kx_k$$

se dice que es la *media ponderada* de los valores $x_1, x_2, \ldots, x_k$; siendo w_i el peso de x_i. Por ejemplo, supongamos que $k = 2$. Si $w_1 = w_2 = 1/2$, la media ponderada

$$w_1x_1 + w_2x_2 = \frac{1}{2}x_1 + \frac{1}{2}x_2$$

coincide exactamente con la media ordinaria de x_1 y x_2. Si, por otra parte, $w_1 = 2/3$ y $w_2 = 1/3$, la media ponderada resultante

$$w_1x_1 + w_2x_2 = \frac{2}{3}x_1 + \frac{1}{3}x_2$$

asigna a x_1 un peso que es el doble del asignado a x_2.

Si se escribe la ecuación (3.1) en la forma

$$\bar{x} = \frac{f_1}{n}x_1 + \frac{f_2}{n}x_2 + \cdots + \frac{f_k}{n}x_k$$

se ve que la media muestral $\bar{x}$ es una media ponderada del conjunto de valores distintos. Los pesos dados al valor distinto x_i es f_i/n, la proporción de valores iguales a x_i. Así, por ejemplo, en el ejemplo 3.3 se podría escribir que

$$\bar{x} = \frac{2}{6} \times 3 + \frac{1}{6} \times 4 + \frac{3}{6} \times 5 = \frac{25}{6}$$

Ejemplo 3.4 En un artículo titulado "Los efectos del uso del casco sobre la gravedad de los daños craneales producidos en los accidentes de moto", publicado en el *Journal of the American Statistical Association* (1992, p. 48-56), A. Weiss analizó una muestra de 770 accidentes de moto similares ocurridos en el área de Los Ángeles en 1976 y 1977. Cada accidente se clasificó según la gravedad del daño sufrido por el conductor. La clasificación utilizada fue la siguiente:

Clasificación del accidente	Interpretación
0	Sin daño craneal
1	Daño menor
2	Daño moderado
3	Grave, sin peligro de muerte
4	Grave, con peligro de muerte
5	Crítico, supervivencia incierta en el momento del accidente
6	Fatal

En 331 de los accidentes el conductor llevaba casco, mientras que en los restantes 439 accidentes no fue así. La siguiente tabla muestra las frecuencias de las distintas gravedades de los accidentes en los que el conductor llevaba puesto el casco y en los que no lo llevaba.

Clasificación	Frecuencia entre los conductores con casco	Frecuencia entre los conductores sin casco
0	248	227
1	58	135
2	11	33
3	3	14
4	2	3
5	8	21
6	1	6
	331	439

Encuentre la media muestral de las clasificaciones de gravedad para los conductores que llevaban casco y para los que no lo llevaban.

Solución La media muestral para los conductores que llevaban casco es

$$\bar{x} = \frac{0\cdot 248 + 1\cdot 58 + 2\cdot 11 + 3\cdot 3 + 4\cdot 2 + 5\cdot 8 + 6\cdot 1}{331} = \frac{143}{331} = 0{,}432$$

La media muestral para aquellos que no llevaban casco es

$$\bar{x} = \frac{0\cdot 227 + 1\cdot 135 + 2\cdot 33 + 3\cdot 14 + 4\cdot 3 + 5\cdot 21 + 6\cdot 6}{439} = \frac{396}{439} = 0{,}902$$

Por consiguiente, los datos indican que aquellos motoristas que llevaban el casco sufrieron, como media, daños craneales menos graves que aquellos que no lo llevaban. ■

3.2.1 Desviaciones

Supongamos de nuevo que los datos muestrales consisten en los n valores $x_1, x_2, \ldots, x_n$, y que $\bar{x} = \sum_{i=1}^{n} x_i/n$ es la media muestral. Las diferencias entre cada uno de los valores y la media muestral se denominan *desviaciones*.

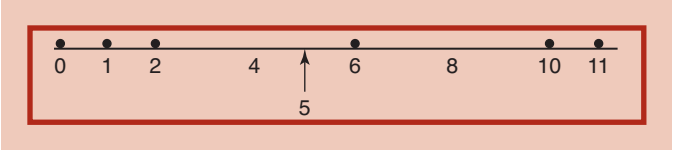


Figura 3.1 El centro de gravedad de 0, 1, 2, 6, 10, 11 es $(0 + 1 + 2 + 6 + 10 + 11)/6 = 30/6 = 5$.

Definición

Las *desviaciones* son las diferencias entre los valores de datos y la media muestral. El valor de la i-ésima desviación es $x_i - \bar{x}$.

Una identidad que puede resultar útil es que la suma de todas las desviaciones debe ser igual a 0. Es decir,

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

La certeza de esta igualdad se puede comprobar como sigue:

$$\begin{aligned} \sum_{i=1}^{n} (x_i - \bar{x}) &= \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} \\ &= n\bar{x} - n\bar{x} \\ &= 0 \end{aligned}$$

Esta igualdad establece que la suma de todas las desviaciones positivas de la media muestral debe compensar exactamente la suma de todas las desviaciones negativas. En términos físicos, esto significa que si se colocan n pesos de igual masa en una varilla (sin peso) en los puntos x_i, $i = 1, \ldots, n$, $\bar{x}$ es el punto en el que la varilla se mantiene en equilibrio. Este centro de equilibrio se conoce con el nombre de *centro de gravedad* (figura 3.1).

Perspectiva histórica

En los primeros viajes marinos era bastante común que gran parte de la carga de un barco resultara dañada debido a las tormentas. Para compensar esta pérdida potencial, existía un acuerdo estándar, en el sentido de que todos aquellos que tenían mercancía a bordo deberían contribuir a pagar por el valor de los artículos perdidos o dañados. La cantidad que se reclamaba a cada uno de ellos se denominaba *havaria*, y de esta palabra latina se deriva el término inglés *average* (media, en español). [De hecho, si existían n personas que transportaban mercancías y las pérdidas sufridas por cada una de ellas fueran $x_1, \ldots, x_n$, la pérdida total sería $x_1 + \cdots + x_n$ y la havaria de cada uno se fijaba en $(x_1 + \cdots + x_n)/n$.]

Ejemplo 3.5 Con los datos del ejemplo 3.1, las desviaciones a la media muestral, 28,48, son

$$x_1 - \bar{x} = 28{,}2 - 28{,}48 = -0{,}28$$

$$x_2 - \bar{x} = 28{,}3 - 28{,}48 = -0{,}18$$

$$x_3 - \bar{x} = 28{,}4 - 28{,}48 = -0{,}08$$

$$x_4 - \bar{x} = 28{,}5 - 28{,}48 = -0{,}02$$

$$x_5 - \bar{x} = 29{,}0 - 28{,}48 = -0{,}52$$

Como comprobación, observe que la suma de las desviaciones es

$$-0{,}28 - 0{,}18 - 0{,}08 + 0{,}02 + 0{,}52 = 0$$ ■

Problemas

1. Los siguientes datos representan las calificaciones en un examen de Estadística para una muestra de estudiantes:

   87, 63, 91, 72, 80, 77, 93, 69, 75, 79, 70, 83, 94, 75, 88

   ¿Cuál es la media muestral?

2. Los siguientes datos (procedentes del Departamento de Agricultura, *Consumo de alimentos, precios y gastos*) muestran el consumo de queso (en libras) per cápita en Estados Unidos durante una muestra de años.

Año	1965	1975	1985	1995	2001
Consumo per cápita	10,0	14,8	23,4	26,4	30,1

   Encuentre la media muestral de los datos anteriores.

3. Los datos siguientes muestran los promedios anuales de lluvia caída (en pulgadas) y de días con precipitación en una muestra de ciudades.

Ciudad	Promedio de lluvia	Promedio de días con precipitación
Albany, NY	35,74	134
Baltimore, MD	31,50	83
Casper, WY	11,43	95
Denver, CO	15,31	88
Fargo, ND	19,59	100
Houston, TX	44,76	105
Knoxville, TN	47,29	127
Los Angeles, CA	12,08	36
Miami, FL	57,55	129
New Orleans, LA	59,74	114
Pittsburgh, PA	36,30	154
San Antonio, TX	29,13	81
Wichita, KS	28,61	85

Fuente: Administración Oceánica y Atmosférica Nacional.

(a) Encuentre la media muestral de los promedios de lluvia en pulgadas.

(b) Encuentre la media muestral de los promedios de los días con precipitación.

4. Considere cinco números y suponga que la media de los cuatro primeros es 14.

   (a) Si el quinto número es 24, ¿cuál es la media de los cinco números?

   (b) Si la media de los cinco números es 24, ¿cuál es el valor del quinto número?

5. Los siguientes datos, sacados del *Resumen estadístico de Estados Unidos,* 1993, muestra el número de policías fallecidos en actos de servicio en Estados Unidos durante los años comprendidos entre 1979 y 1990. Encuentre la media muestral de estos datos.

   164, 165, 157, 164, 152, 147, 148, 131, 147, 155, 145, 132

6. Suponga que la media muestral de un conjunto de 10 datos puntuales es $\bar{x} = 20$.

   (a) Si se descubre que se ha leído incorrectamente un dato con valor 15 y que se le ha dado el valor 13, ¿cuál será el valor revisado de la media muestral?

   (b) Si existiera un dato adicional con valor 22, ¿aumentaría o disminuiría el valor de $\bar{x}$?

   (c) Con los datos originales [y no con los datos revisados en el apartado (a)], ¿cuál sería el nuevo valor de $\bar{x}$ del apartado (b)?

7. La tabla siguiente lista la cantidad anual de casos de tétano que han sido notificados en Estados Unidos para una muestra de años. Calcule la media muestral.

Año	1970	1975	1980	1982	1984	1985	1987	2001
Número de casos	148	102	95	88	74	83	48	62

Fuente: Centro de control de enfermedades de Estados Unidos. *Sumario de enfermedades notificables, morbosidad y mortalidad.*

8. El siguiente gráfico de tallos y hojas refleja los puntos obtenidos por el autor de este texto en 15 juegos de bolos. Calcule la media muestral.

18	2, 4, 7
17	0
16	1, 9
15	2, 2, 4, 8, 8
14	
13	2, 1, 5, 5

9. Encuentre la media muestral de este conjunto de datos:

$$1, 2, 4, 7, 10, 12$$

Calcule, ahora, las medias de los conjuntos de datos

$$3, 6, 12, 21, 30, 36 \qquad \text{y} \qquad 6, 7, 9, 12, 15, 17$$

10. Suponga que $\bar{x}$ es la media muestral de un conjunto de datos compuesto por los valores $x_1, \ldots, x_n$. Si los datos se transforman de acuerdo con la expresión

$$y_i = ax_i + b \qquad i = 1, \ldots, n$$

¿Cuál es la media muestral del conjunto de datos $y_1, \ldots, y_n$? (En la expresión anterior, a y b son constantes dadas.)

11. Los datos siguientes reflejan el número total de incendios en Ontario (Canadá), ocurridos en los sucesivos meses del año 2002.

$$6, 13, 5, 7, 7, 3, 7, 2, 5, 6, 9, 8$$

Encuentre la media muestral de este conjunto de datos.

12. El siguiente conjunto de datos muestra el número total de coches vendidos en Estados Unidos en una muestra de años. Los datos están dados en unidades de miles de coches. Encuentre la media muestral del número de coches vendidos anualmente en dichos años.

Año	1980	1985	1990	1995	2000	2002
Número de ventas	8010	11 653	9783	11 985	12 832	12 326

Fuente: Resumen estadístico de Estados Unidos, 1990.

13. La mitad de los valores de una muestra son iguales a 10, y los de la otra mitad son todos iguales a 20. ¿Cuál es la media muestral?

14. La siguiente tabla de frecuencias refleja las edades de los componentes de una joven orquesta sinfónica.

Edades	Frecuencias
16	9
17	12
18	15
19	10
20	8

Encuentre la media muestral de las edades dadas.

15. En una muestra de datos, la mitad de los valores son iguales a 10, una sexta parte son iguales a 20 y una tercera parte son iguales a 30. ¿Cuál es la media muestral?

16. Existen dos entradas a un aparcamiento. El estudiante 1 contabiliza el número de coches que pasan diariamente a través de la entrada 1, y el estudiante 2 hace lo mismo en la entrada 2. A lo largo de 30 días, los datos del estudiante 1 tienen una media muestral de 122, mientras la media muestral de los datos del estudiante 2 es igual a 160. Sobre los 30 días citados, ¿cuál fue el número medio de coches que entraron en el aparcamiento?

17. Una compañía tiene dos plantas de producción. El salario medio de una muestra de 30 ingenieros de la planta 1 fue de 33 600 dólares, mientras que el salario medio de una muestra de 20 ingenieros de la planta 2 resultó ser de 42 400 dólares. ¿Cuál es el salario medio muestral de de los 50 ingenieros seleccionados?

18. Supongamos que se dispone de dos muestras distintas, de tamaños n_1 y n_2. Si la media muestral de la primera muestra es $\bar{x}_1$ y la de la segunda muestra es $\bar{x}_2$, ¿cuál es la media de la muestra conjunta, de tamaño $n_1 + n_2$?

19. Encuentre las desviaciones para cada uno de los tres conjuntos de datos del problema 9, y contraste la veracidad de sus respuestas mediante la comprobación de que, en cada caso, la suma de las desviaciones es 0.

20. Calcule las desviaciones de los datos del problema 14 y compruebe que su suma es 0.

3.3 Mediana muestral

Los siguientes datos representan el número de semanas que, tras completar un curso para aprender a conducir, tardó cada miembro de una muestra de siete personas en obtener su carné de conducir:

$$2, 110, 5, 7, 6, 7, 3$$

La media muestral de este conjunto de datos es $\bar{x} = 140/7 = 20$; así pues, seis de los siete valores de datos son menores que la media muestral, mientras que el séptimo valor es muy

superior a ésta. Lo que apunta una debilidad de la media muestral como indicador del centro de un conjunto de datos: a saber, su valor se encuentra muy afectado por los valores de datos extremos.

Un estadístico que se utiliza también para representar el centro de un conjunto de datos es la *mediana muestral*, definida como el valor medio cuando los datos están ordenados de menor a mayor. La mediana muestral será denotada por m.

Definición

Ordene los valores de datos de menor a mayor. Si el número de datos es impar, la *mediana muestral* coincide con el valor que se encuentra en la posición central en la lista ordenada; si el número de datos es par, la *mediana muestral* es la media de los dos valores que ocupan las posiciones centrales.

De esta definición se deduce que, si existen tres valores de datos, la mediana muestral coincide con el segundo valor más pequeño; mientras que, si existen cuatro valores, coincide con la media de los valores más pequeños segundo y tercero.

Ejemplo 3.6 Los siguientes datos representan el número de semanas que siete individuos tardaron en obtener su carné de conducir. Encuentre la mediana muestral.

$$2, 110, 5, 7, 6, 7, 3$$

Solución Ordenemos primero los datos en orden creciente.

$$2, 3, 5, 6, 7, 7, 110$$

Puesto que el tamaño de la muestra es 7, la mediana muestral será el cuarto valor más pequeño. Esto es, la mediana muestral del número de semanas que se tardó en obtener el carné de conducir es $m = 6$ semanas. ■

Ejemplo 3.7 Los siguientes datos representan el número de días que a seis individuos les costó dejar de fumar tras completar un cursillo diseñado para este propósito.

$$1, 2, 3, 5, 8, 100$$

¿Cuál es la mediana muestral?

Solución Puesto que el tamaño muestral es 6, la mediana muestral es la media de los dos valores centrales una vez ordenados; así pues,

$$m = \frac{3 + 5}{2} = 4$$

Es decir, la mediana muestral es de 4 días. ■

En general, para un conjunto de datos de n valores, la mediana muestral coincide con el $[(n + 1)/2]$ menor valor ordenado, cuando n es impar, y coincide con la media de los valores ordenados que ocupan las posiciones $(n/2)$ y $(n/2 + 1)$, cuando n es par.

Tanto la media muestral como la mediana muestral son estadísticos útiles para describir la tendencia central de un conjunto de datos. La media muestral, siendo la media aritmética, utiliza todos los valores de datos. La mediana muestral, puesto que sólo utiliza un único valor central o bien un par de valores centrales, no se ve afectada por los valores extremos.

Ejemplo 3.8 Los datos siguientes proporcionan los nombres de los máximos encestadores individuales de la Asociación de Baloncesto Nacional (NBA) junto con su promedio de puntos por partido en las temporadas comprendidas entre 1953 y 1991.

Temporada	Jugador, equipo	Promedio de puntos
1953–54	Neil Johnston, Philadelphia Warriors	24,4
1954–55	Neil Johnston, Philadelphia Warriors	22,7
1955–56	Bob Pettit, St. Louis Hawks	25,7
1956–57	Paul Arizin, Philadelphia Warriors	25,6
1957–58	George Yardley, Detroit Pistons	27,8
1958–59	Bob Pettit, St. Louis Hawks	29,2
1959–60	Wilt Chamberlain, Philadelphia Warriors	37,6
1960–61	Wilt Chamberlain, Philadelphia Warriors	38,4
1961–62	Wilt Chamberlain, Philadelphia Warriors	50,4
1962–63	Wilt Chamberlain, San Francisco Warriors	44,8
1963–64	Wilt Chamberlain, San Francisco Warriors	36,9
1964–65	Wilt Chamberlain, San Francisco Warriors–Phila. 76ers	34,7
1965–66	Wilt Chamberlain, Philadelphia 76ers	33,5
1966–67	Rick Barry, San Francisco Warriors	35,6
1967–68	Dave Bing, Detroit Pistons	27,1
1968–69	Elvin Hayes, San Diego Rockets	28,4
1969–70	Jerry West, Los Angeles Lakers	31,2
1970–71	Lew Alcindor, Milwaukee Bucks	31,7
1971–72	Kareem Abdul-Jabbar, Milwaukee Bucks	34,8
1972–73	Nate Archibald, Kansas City–Omaha Kings	34,0
1973–74	Bob McAdoo, Buffalo Braves	30,8
1974–75	Bob McAdoo, Buffalo Braves	34,5
1975–76	Bob McAdoo, Buffalo Braves	31,1
1976–77	Pete Maravich, New Orleans Jazz	31,1
1977–78	George Gervin, San Antonio Spurs	27,2
1978–79	George Gervin, San Antonio Spurs	29,6
1979–80	George Gervin, San Antonio Spurs	33,1
1980–81	Adrian Dantley, Utah Jazz	30,7
1981–82	George Gervin, San Antonio Spurs	32,3
1982–83	Alex English, Denver Nuggets	28,4
1983–84	Adrian Dantley, Utah Jazz	30,6
1984–85	Bernard King, New York Knicks	32,9
1985–86	Dominique Wilkins, Atlanta Hawks	30,3

Temporada	Jugador, equipo	Promedio de puntos
1986–87	Michael Jordan, Chicago Bulls	37,1
1987–88	Michael Jordan, Chicago Bulls	35,0
1988–89	Michael Jordan, Chicago Bulls	32,5
1989–90	Michael Jordan, Chicago Bulls	33,6
1990–91	Michael Jordan, Chicago Bulls	31,5
1991–92	Michael Jordan, Chicago Bulls	30,1

(a) Encuentre la mediana muestral del promedio de puntos.

(b) Calcule la media muestral del promedio de puntos.

Elimine las temporadas que comienzan en 1961 y en1962, en las que Wilt Chamberlain tuvo un promedio de 50,4 y 44,8 puntos por partido, respectivamente, y encuentre

(c) la mediana muestral

(d) la media muestral

Solución

(a) Puesto que existen 39 valores de datos, la mediana muestral coincide con el 20º valor menor. Existen 11 valores entre 20 y 29, por tanto, la mediana será el noveno valor menor cuando se eliminen todos los promedios inferiores a 30. Si se ordenan los restantes valores se obtiene

$$30{,}1,\ 30{,}3,\ 30{,}6,\ 30{,}7,\ 30{,}8,\ 31{,}1,\ 31{,}1,\ 31{,}2,\ 31{,}5,\ \ldots$$

En consecuencia, la mediana muestral es

$$m = 31{,}5$$

(b) La suma de los 39 valores es 1256,9 y, por tanto, la media muestral es

$$\bar{x} = \frac{1256{,}9}{39} \approx 32{,}228$$

Perspectiva histórica

El matemático holandés Christian Huygens fue uno de los primeros científicos que se dedicaron a la Teoría de la Probabilidad. En 1669, su hermano Ludwig, después de estudiar las tablas de mortalidad de la época, escribió a su famoso hermano mayor: “He estado confeccionando una tabla que muestra cuánto tiempo puede vivir la gente… ¡Vivir bien! De acuerdo con mis cálculos, tú deberías vivir unos $56\frac{1}{2}$ años, y yo 55.” Christian, intrigado, analizó las tablas de mortalidad, pero llegó a unos estimadores, respecto a los años que ambos deberían vivir, distintos de los de su hermano. ¡Ludwig basó sus estimadores en la mediana muestral, mientras que Christian se basó en la media muestral!

(c) Si se eliminan los dos años especificados, la mediana es el 19º valor menor de los 37 valores restantes. A partir de la ordenación dada en (a), que comienza en el 12º valor menor, se obtiene que la mediana muestral es ahora

$$m = 31{,}2$$

(d) Si se eliminan los dos años citados, la suma de todos los valores de datos restantes se reduce a

$$1256{,}9 - 50{,}4 - 44{,}8 = 1161{,}7$$

Por tanto, la media muestral es ahora

$$\bar{x} = \frac{1161{,}7}{37} = 31{,}397$$

Así pues, se ve que eliminar los dos mayores valores del conjunto de datos tiene un efecto relativamente pequeño sobre la mediana, y la reduce de 31,50 a 31,20; mientras que su efecto sobre la media es mucho mayor, pues la reduce de 32,23 a 31,40. ■

Para los conjuntos de datos aproximadamente simétricos sobre su valor central, la media muestral y la mediana muestral tienen valores próximos. Por ejemplo, los datos

$$4, 6, 8, 8, 9, 12, 15, 17, 19, 20, 22$$

son grosso modo simétricos alrededor del valor 12, que es su mediana muestral. La media muestral es $\bar{x} = 140/11 = 12{,}73$, que se encuentra próxima a 12.

La respuesta a la pregunta sobre cuál de los dos estadísticos sumariales es más informativo depende de qué es lo que se pretende conocer del conjunto de datos. Por ejemplo, si el gobierno establece un impuesto sobre la renta con tarifa plana (proporcional) y se pretende averiguar qué recaudación cabe esperar, la renta media de los ciudadanos será más interesante que la mediana (¿por qué?). Por el contrario, si el gobierno estuviera interesado en determinar un valor central de la cantidad de renta que los ciudadanos dedican a la vivienda, la mediana muestral podría ser más informativa (¿por qué?).

Aunque es interesante analizar si la media muestral o la mediana muestral es mas informativa, en una situación concreta, obseve que no debemos restringir nuestro conocimiento a sólo una de dichas magnitudes. Ambas son importantes y, por tanto, las dos se han de calcular cuando se está sintetizando un conjunto de datos.

Problemas

1. Los siguientes datos exponen las distancias que se recorren en una muestra de cursos de golf municipales.

7040, 6620, 6050, 6300, 7170, 5990, 6330, 6780, 6540, 6690, 6200, 6830

(a) Encuentre la mediana muestral.

(b) Encuentre la media muestral.

2. (a) Determine la mediana muestral del conjunto de datos.

$$14, 22, 8, 19, 15, 7, 8, 13, 20, 22, 24, 25, 11, 9, 14$$

(b) Incremente cada valor de (a) en 5 unidades, y encuentre la nueva mediana muestral.

(c) Multiplique por 3 cada valor de (a), y encuentre la nueva mediana muestral.

3. Si la mediana de un conjunto de datos x_i, $i = 1, \ldots, n$, es 10, ¿cuál es la mediana del conjunto de datos $2x_i + 3$, $i = 1, \ldots, n$?

4. Los siguientes datos reflejan las velocidades de 40 coches, medidas por radar en una calle de cierta ciudad.

22, 26, 31, 38, 27, 29, 33, 40, 36, 27, 25, 42, 28, 19, 28, 26, 33, 26, 37, 22,
31, 30, 44, 29, 25, 17, 46, 28, 31, 29, 40, 38, 26, 43, 45, 21, 29, 36, 33, 30

Encuentre la mediana muestral.

5. Las tablas siguientes muestran las tasas de suicidio, de hombres y mujeres, por 100 000 individuos para un conjunto de países.

Tasas de suicidios por 100 000 individuos

Sexo	Estados Unidos	Australia	Austria	Canadá	Dinamarca	Francia
Mujeres	5,4	5,1	15,8	5,4	20,6	12,7
Hombres	19,7	18,2	42,1	20,5	35,1	33,1

Sexo	Italia	Japón	Holanda	Polonia	Suecia	Reino Unido	Alemania
Mujeres	4,3	14,9	8,1	4,4	11,5	5,7	12,0
Hombres	11,0	27,8	14,6	22,0	25,0	12,1	26,6

Fuente: Organización Mundial de la Salud, *Estadística sobre la Salud Mundial.*

(a) Encuentre la mediana muestral para las tasas de suicidio de los hombres.

(b) Encuentre la mediana muestral para las tasas de suicidio de las mujeres.

(c) Encuentre la media muestral para las tasas de suicidio de los hombres.

(d) Encuentre la media muestral para las tasas de suicidio de las mujeres.

6. Obtenga la mediana muestral del número medio anual de días de lluvia para las ciudades incluidas en el problema 3 de la sección 3.2.

7. Encuentre la mediana muestral del número medio anual de pulgadas de precipitación para las ciudades incluidas en el problema 3 de la sección 3.2.

8. Busque la mediana muestral de los datos presentados en el problema 8 de la sección 2.3.

9. Utilice la tabla sobre tasas de mortalidad que antecede al problema 9 de la sección 2.3 para calcular la mediana de las tasas de mortalidad debidas a:

   (a) caídas

   (b) envenenamientos

   (c) ahogamientos

10. La mediana muestral de 10 valores distintos es 5. Deduzca qué se podría decir acerca de la nueva mediana muestral si:

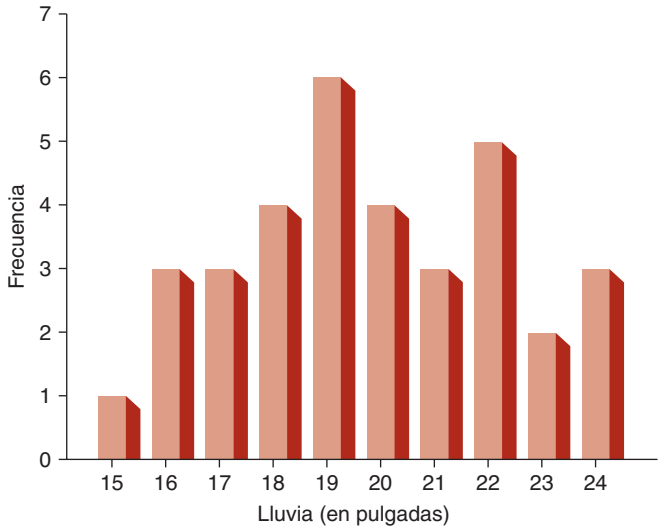
    (a) Al conjunto de datos se le añade un nuevo dato con valor 7.

    (b) Se añaden dos nuevos datos con valores 3 y 42.

11. El histograma de la figura de abajo representa las lluvias anuales, en pulgadas, caídas en una ciudad occidental durante los últimos 34 años. Puesto que los datos originales no son recuperables a partir del histograma, no se puede utilizar éste para calcular el valor de la media muestral y la mediana muestral. Aún así, basándonos en este histograma, diga cuál es el mayor valor posible de:

    (a) la media muestral

    (b) la mediana muestral



Diga, cuál es el menor valor posible de:

(c) la media muestral

(d) la mediana muestral

(e) Los datos reales se muestran a continuación:

15,2, 16,1, 16,5, 16,7, 17,2, 17,5, 17,7, 18,3, 18,6, 18,8, 18,9, 19,1,
19,2, 19,2, 19,6, 19,8, 19,9, 20,2, 20,3, 20,3, 20,8, 21,1, 21,4, 21,7,
22,2, 22,5, 22,5, 22,7, 22,9, 23,3, 23,6, 24,1, 24,5, 24,9

Determine la media muestral y la mediana muestral, y compruebe si concuerdan con sus respuestas anteriores.

12. Los datos siguientes muestran las temperaturas máxima y mínima (en grados Fahrenheit) del 4 de julio de 1993 en varias ciudades.

Ciudad	Temperaturas máxima/mínima del 4 de julio de 1993
Atlanta	96/75
Boise	75/53
Cleveland	90/68
Jacksonville	95/75
Norfolk	89/73
Providence	89/68
Rochester	85/59
Seattle	68/55
Toledo	93/71
Wilmington	95/71

Fuente: Periódico *New York Times*, 5 de julio de 1993.

(a) Encuentre la mediana muestral de las temperaturas máximas.

(b) Encuentre la mediana muestral de las temperaturas mínimas.

(c) Encuentre la mediana muestral de las diferencias entre las temperaturas máxima y mínima.

13. Con los datos del ejemplo 3.4, calcule las medianas muestrales de las lesiones craneales graves sufridas por los conductores de moto que llevaban puesto el casco y por los que no lo llevaban.

13. En las situaciones siguientes, ¿cuál de los estadísticos media muestral o mediana muestral piensa que es más informativo?

(a) Para analizar si se debe cerrar una línea de autobús entre Rochester y Nueva York, un ejecutivo ha recopilado el número de viajeros en una muestra de días.

(b) Para comparar a los estudiantes universitarios actuales con los de años anteriores, se consultan muestras de las calificaciones obtenidas en los exámenes de acceso a la universidad durante varios años.

(c) El abogado defensor de un proceso judicial con jurado popular está analizando las puntuaciones de un test de inteligencia (IQ) obtenidas por los miembros del jurado.

(d) Usted ha comprado su casa hace 6 años en una pequeña comunidad por un precio de 105 000 dólares, que coincidía con el precio medio y mediano de todas las casas que se vendieron aquel año en dicha comunidad. Sin embargo, en los dos últimos

años, se han construido varias casas nuevas mucho más caras que las anteriores. Para obtener una idea del valor actual de su casa, usted decide analizar los precios de venta de las casas vendidas recientemente en su comunidad.

15. Las mujeres suponen los siguientes porcentajes de fuerza laboral en las ocupaciones que se listan a continuación.

Ocupación	Porcentaje de mujeres	Ocupación	Porcentaje de mujeres
Ejecutivas de empresa	36,8	Médicos	17,6
Enfermeras	94,3	Abogadas	18,0
Supervisoras de ventas	30,5	Profesoras de enseñanza básica	85,2
Vendedoras	68,6	Empleadas de correos	43,5
Bomberas	1,9	Policías	10,9
Empleadas de la limpieza	41,5	Supervisoras de construcción	1,6
Trabajadoras de la construcción	2,8	Conductoras de camiones	2,1

Encuentre para estos porcentajes:

(a) la media muestral

(b) la mediana muestral

Adicionalmente, resulta que las mujeres representan el 44,4% de la fuerza laboral total en todas las ocupaciones anteriores. ¿Esto resulta coherente con sus respuestas a los apartados (a) y (b)? Explique por qué.

16. Con los datos relativos a los 30 primeros estudiantes del Apéndice A, encuentre la mediana muestral y la media muestral de:

(a) los pesos

(b) los niveles de colesterol

(c) las presiones sanguíneas

17. La tabla siguiente muestra las edades medianas de hombres y mujeres en el momento de contraer su primer matrimonio, correspondientes a las bodas celebradas entre los años 1992 y 2002.

(a) Encuentre la mediana muestral de la edad mediana de los hombres.

(b) Encuentre la mediana muestral de la edad mediana de las mujeres.

Edad mediana en el primer matrimonio, en Estados Unidos

Año	Hombres	Mujeres	Año	Hombres	Mujeres
2002	26,9	25,3	1996	27,1	24,8
2001	26,9	25,1	1995	26,9	24,5
2000	26,8	25,1	1994	26,7	24,5
1999	26,9	25,1	1993	26,5	24,5
1998	26,7	25,0	1992	26,5	24,4
1997	26,8	25,0			

3.3.1 Percentiles muestrales

La mediana muestral es un caso particular de los estadísticos conocidos como *percentiles muestrales de orden 100p por ciento*, donde p puede ser cualquier valor comprendido entre 0 y 1. Grosso modo, el percentil muestral de orden $100p$ por ciento es aquel valor que verifica que el $100p$ por ciento de los valores de los datos son menores que él y que el $100(1—p)\%$ de los valores de los datos son mayores que él.

Definición

El *percentil muestral de orden 100p por ciento* es aquel valor de dato que tiene la propiedad de que al menos el $100p$ por ciento de los valores de datos son menores o iguales que él y que al menos el $100(1—p)$ por ciento de los valores de datos son mayores o iguales que él. Si existen dos valores de datos que cumplen las condiciones anteriores, el percentil muestral de orden $100p$ por ciento se define como la media aritmética de ambos valores de datos.

Observe que la mediana muestral se corresponde con el percentil muestral de orden 50%. Es decir, coincide con el percentil muestral de orden $100p$ por ciento cuando $p = 0{,}50$.

Supongamos que se han ordenado de menor a mayor todos los valores de datos de una muestra de tamaño n. Para determinar el percentil muestral de orden $100p$ por ciento se debe encontrar aquel valor que verifica que:

1. Al menos np valores de datos son menores o iguales que él.
2. Al menos $n(1—p)$ valores de datos son mayores o iguales que él.

Ahora bien, si np no es un entero, el único valor de dato que cumple ambos puntos es aquel cuya posición de orden coincide con el primer entero superior a np. Por ejemplo, supongamos que se quiere determinar el percentil muestral de orden 90% en una muestra de tamaño $n = 12$. Puesto que $p = 0{,}9$, se tiene que $np = 10{,}8$ y $n(1—p) = 1{,}2$. Así pues, se estarán buscando aquellos valores para los que:

1. Al menos 10,8 valores de datos sean menores o iguales que él (por consiguiente, el valor de datos debe estar en la posición de orden 11 o mayor).
2. Al menos 1,2 valores de datos sean mayores o iguales que él (por tanto, debe ocupar la posición de orden 11 o menor).

Evidentemente, el único valor de dato que cumple ambos puntos es aquél que ocupa la posición de orden 11, y, en consecuencia, éste será el percentil muestral de orden 90%.

Por otro lado, si np es un entero, tanto el valor de dato que ocupa la np posición de orden como el valor de dato que ocupa la posición de orden $np + 1$ cumplen las condiciones de las definiciones 1. y 2.; en este caso, el percentil muestral de orden $100p$ por ciento será igual a la media aritmética de los dos valores de datos anteriores. Por ejemplo, supongamos que se desea encontrar el percentil muestral de orden 95% en un conjunto de datos con $n = 20$ valores. En este caso, tanto el 19º valor como el 20º valor (los dos valores mayores) serán mayores o iguales que al menos $np = 20(0{,}95) = 19$ de los valores de datos, y serán menores o iguales que al menos $n(1—p) = 1$ de dichos valores. El percentil muestral de orden 95% será, pues, la media aritmética de los valores que ocupan las posiciones de orden 19 y 20 (es decir, los dos mayores).

Resumiendo, se ha demostrado lo siguiente.

Para encontrar el percentil muestral de orden 100*p*% de un conjunto de datos de tamaño *n*

1. Ordene los datos en sentido creciente.
2. Si np no es un entero, determine el menor entero mayor que np. El valor de dato que ocupa la posición de orden igual a este último entero será el percentil muestral de orden $100p$ por ciento.
3. Si np es un entero, el percentil muestral de orden $100p$ por ciento coincidirá con la media aritmética de los valores que ocupan las posiciones de orden np y $np + 1$.

Ejemplo 3.9 ¿Cómo se calcula el percentil muestral de orden 90% cuando el tamaño muestral es (a) 8, (b) 16 y (c) 100?

Solución

(a) Puesto que $0{,}9 \times 8 = 7{,}2$ no es un entero, si se ordenan los datos de menor a mayor, el percentil muestral de orden 90% coincidirá con el octavo valor menor (es decir, el valor mayor).

(b) Puesto que $0{,}9 \times 16 = 14{,}4$ no es un entero, el percentil muestral de orden 90% será el 15º valor menor.

(c) Puesto que $0{,}9 \times 100 = 90$ es un entero, el percentil muestral de orden 90% coincidirá con la media aritmética de los valores que ocupan las posiciones 90 y 91 una vez que los datos han sido ordenados de menor a mayor. ■

Ejemplo 3.10 La tabla 3.1 lista las primeras 20 universidades de Estados Unidos en una clasificación basada en los activos que han generado. Utilice estos datos para encontrar:

(a) el percentil muestral de orden 90%

(b) el percentil muestral de orden 20%

Solución

(a) Puesto que el tamaño muestral es 20 y $20 \times 0{,}9 = 18$, el percentil muestral de orden 90% coincide con la media aritmética entre los valores más pequeños 18º y 19º o, equivalentemente, la media aritmética entre los valores más grandes 2º y 3º. De donde:

$$\text{Percentil muestral de orden 90\%} = \frac{10\,523\,600 + 8\,630\,679}{2} = 9\,977\,140$$

Es decir, el percentil muestral de orden 90% para este conjunto de datos es aproximadamente igual a 9,98 miles de millones de dólares.

Tabla 3.1 Las 20 universidades más altas en la clasificación de becas generadas, 2002*

Universidad	Activos†	Universidad	Activos†
1. Harvard University	17 169 757 $	11. Washington University	3 517 104 $
2. Yale University	10 523 600	12. University of Pennsylvania	3 393 297
3. University of Texas System	8 630 679	13. University of Michigan	3 375 689
4. Princeton University	8 319 600	14. University of Chicago	3 255 368
5. Stanford University	7 613 000	15. Northwestern University	3 022 733
6. Massachusetts Institute of Technology	5 359 423	16. Rice University	2 939 804
		17. Duke University	2 927 478
7. Emory University	4 551 873	18. Cornell University	2 853 742
8. Columbia University	4 208 373	19. University of Notre Dame	2 554 004
9. University of California	4 199 067	20. Dartmouth College	2 186 610
10. The Texas A&M University System and Foundations	3 743 442		

Observación: Valor de mercado de los activos generados, excluyendo las donaciones privadas y el capital de trabajo.
* Con fecha de 30 de junio de 2002.
† En miles.
Fuente: Asociación Nacional de Agentes de Negocios Universitarios (NACUBO).

(b) Puesto que $20 \times 0{,}2 = 4$, el percentil muestral de orden 20% es el promedio entre los valores menores 4º y 5º, se obtiene el resultado:

$$(\text{Percentil muestral de orden } 20\%) = \frac{2\,927\,478 + 2\,939\,804}{2} = 2\,823\,641 \quad \blacksquare$$

Los percentiles muestrales de órdenes 25, 50 y 75% se conocen como *cuartiles*.

Definición

El percentil muestral de orden 25% se llama *primer cuartil*. El percentil muestral de orden 50% se denomina *mediana* o *segundo cuartil*. El percentil muestral de orden 75% se llama *tercer cuartil*.

Los cuartiles dividen el conjunto de datos en cuatro partes, de forma que, aproximadamente, un 25% de los valores de datos se encuentran por debajo del primer cuartil, otro 25% de los valores se encuentra entre el primer y el segundo cuartil, un tercer 25% se encuentra entre el segundo y el tercer cuartil y, por último, el 25% restante de los valores supera al tercer cuartil.

Ejemplo 3.11 Encuentre los cuartiles muestrales para los siguientes 18 valores de datos, que se muestran ordenados y representan las puntuaciones de una liga de bolos.

122, 126, 133, 140, 145, 145, 149, 150, 157, 162, 166, 175, 177, 177, 183, 188, 199, 212

Solución Puesto que $0{,}25 \times 18 = 4{,}5$, el percentil muestral de orden 25% coincide con el quinto valor menor, que es 145.

Dado que $0{,}50 \times 18 = 9$, el segundo cuartil (o mediana muestral) es igual a la media del noveno y décimo valor menor, es decir, su valor es:

$$\frac{157 + 162}{2} = 159{,}5$$

Finalmente, puesto que $0{,}75 \times 18 = 13{,}5$, el tercer cuartil coincide con el 15º valor menor, que es 177. ■

Problemas

1. Se han ordenado 75 valores en sentido creciente. ¿Cómo se determinarían los percentiles muestrales siguientes del conjunto de datos?

   (a) percentil de orden 80%

   (b) percentil de orden 60%

   (c) percentil de orden 30%

2. La siguiente tabla muestra las exportaciones de plátanos, en toneladas métricas, en una selección de países iberoamericanos y caribeños. Encuentre los cuartiles.

Exportaciones de plátanos, año 2000, en toneladas métricas

Ecuador	4 095 191
Costa Rica	2 113 652
Colombia	1 710 949
Guatemala	857 164
Panamá	489 805
Honduras	183 400
México	81 044
Santa Lucía	72 795
Brasil	72 468
Belice	64 400
República Dominicana	62 429
Nicaragua	44 402
San Vicente y las Granadinas	43 810
Jamaica	40 900
Surinam	34 000
Venezuela	33 543
Dominica	29 810
Bolivia	9 377

Exportaciones de plátanos, año 2000, en toneladas métricas (*Continuación*)

Perú	856
Granada	707
Argentina	412
Trinidad y Tobago	87
El Salvador	72
Paraguay	66
Chile	18
Guayana	10
Total	10 041 367

Fuente: FAO.

3. Considere un conjunto de datos con n valores. Diga cómo se calcula el percentil muestral de orden 95% cuando

   (a) $n = 100$

   (b) $n = 101$

La tabla siguiente muestra el número de médicos y dentistas por cada 100 000 habitantes para 12 Estados del occidente medio de Estados Unidos en el año 2000. Los problemas 4 y 5 se basan en ella.

Estado	**Tasa de médicos**	**Tasa de dentistas**
Ohio	188	56
Indiana	146	48
Illinois	206	61
Michigan	177	64
Wisconsin	177	70
Minnesota	207	70
Iowa	141	60
Missouri	186	55
North Dakota	157	55
South Dakota	129	54
Nebraska	162	71
Kansas	166	52

Fuente: Asociación Médica Americana, *Características y Distribuciones Médicas en Estados Unidos.*

4. Encuentre, para las tasas de médicos por cada 100 000 habitantes:

   (a) el percentil muestral de orden 40%

   (b) el percentil muestral de orden 60%

   (c) el percentil muestral de orden 80%

5. Encuentre, para las tasas de dentistas por cada 100 000 habitantes:
   (a) el percentil muestral de orden 90%
   (b) el percentil muestral de orden 50%
   (c) el percentil muestral de orden 10%
6. Supongamos que el percentil muestral de orden 100p por ciento para un conjunto de datos es 120. Si se suma 30 a cada valor de dato, ¿cuál es el nuevo valor del percentil muestral de orden 100p por ciento?
7. Supongamos que el percentil muestral de orden 100p por ciento para un conjunto de datos es 230. Si se multiplica cada valor por una constante positiva c, ¿cuál es el nuevo valor del percentil muestral de orden 100p por ciento?
8. Encuentre el percentil muestral de orden 90% del siguiente conjunto de datos.

   75, 33, 55, 21, 46, 98, 103, 88, 35, 22, 29, 73, 37, 101, 121, 144, 133, 52, 54, 63, 21, 7
9. La tabla siguiente muestra los fallecimientos por accidentes de tráfico (por 100 millones de millas recorridas) en el año 2001 en los 50 Estados y en el distrito de Columbia de Estados Unidos. Encuentre los cuartiles.

Muertes por accidente de tráfico por 100 millones de millas recorridas, 200

Estado	**Tasa**	**Rango de orden**
Estados Unidos	1,51	(X)
Alabama	1,75	16
Alaska	1,80	15
Arizona	2,06	7
Arkansas	2,08	6
California	1,27	37
Colorado	1,71	19
Connecticut	1,01	47
Delaware	1,58	24
District of Columbia	1,81	(X)
Florida	1,93	10
Georgia	1,50	26
Hawaii	1,61	23
Idaho	1,84	13
Illinois	1,37	31
Indiana	1,27	37
Iowa	1,49	27
Kansas	1,75	16
Kentucky	1,83	14
Louisiana	2,32	1
Maine	1,33	34
Maryland	1,27	37
Massachusetts	0,90	50
Michigan	1,34	33

Muertes por accidente de tráfico por 100 millones de millas recorridas, 200 (*Continuación*)

Estado	Tasa	Rango de orden
Minnesota	1,06	46
Mississippi	2,18	4
Missouri	1,62	22
Montana	2,30	2
Nebraska	1,36	32
Nevada	1,71	19
New Hampshire	1,15	44
New Jersey	1,09	45
New Mexico	1,99	9
New York	1,18	43
North Carolina	1,67	21
North Dakota	1,45	29
Ohio	1,29	36
Oklahoma	1,55	25
Oregon	1,42	30
Pennsylvania	1,49	27
Rhode Island	1,01	47
South Carolina	2,27	3
South Dakota	2,00	8
Tennessee	1,85	12
Texas	1,72	18
Utah	1,25	41
Vermont	0,96	49
Virginia	1,27	37
Washington	1,21	42
West Virginia	1,91	11
Wisconsin	1,33	34
Wyoming	2,16	5

Observación: : Cuando dos o más Estados comparten el mismo rango de orden, los siguientes rangos de orden se omiten. Debido al redondeo de datos, varios Estados pueden tener valores idénticos, aunque su rango sea distinto.

10. Los cuartiles de un extenso conjunto de datos son los siguientes:

$$\begin{aligned}\text{Primer cuartil} &= 35\\ \text{Segundo cuartil} &= 47\\ \text{Tercer cuartil} &= 66\end{aligned}$$

(a) Indique un intervalo que contenga aproximadamente un 50% de los datos.

(b) Determine un valor que aproximadamente sea mayor que un 50% de los datos.

(c) Determine un valor para el que aproximadamente un 25% de los datos sean mayores que él.

11. La mediana de un conjunto de datos simétrico es igual a 40 y su tercer cuartil es igual a 55. ¿Cuál es valor del primer cuartil?

3.4 Moda muestral

Otro indicador de la tendencia central es la *moda muestral*, que se define como el valor de dato que aparece con mayor frecuencia en un conjunto de datos.

Ejemplo 3.12 Los siguientes datos se refieren a las tallas de los últimos 8 vestidos vendidos en una boutique de mujeres.

8, 10, 6, 4, 10, 12, 14, 10

¿Cuál es la moda muestral?

Solución La moda muestral es 10, puesto que este es el valor que ocurre con mayor frecuencia. ■

Si no existe un único valor que aparezca con mayor frecuencia en el conjunto de datos, aquellos valores que tengan la máxima frecuencia se denominan *valores modales*. En esta situación se dice que no existe un valor único de la moda muestral.

Ejemplo 3.13 Las edades de 6 niños de una guardería son las siguientes:

2, 5, 3, 5, 2, 4

¿Cuáles son los valores modales de este conjunto de datos?

Solución Puesto que las edades 2 y 5 son las que ocurren con mayor frecuencia, estos dos valores (2 y 5) son los modales. ■

Resulta muy sencillo obtener el valor modal a partir de una tabla de frecuencias, puesto que coincide con aquel valor que tenga mayor frecuencia.

Ejemplo 3.14 La siguiente tabla de frecuencias muestra los valores obtenidos en 30 lanzamientos de un dado.

Valor	Frecuencia
1	6
2	4
3	5
4	8
5	3
6	4

Para estos datos, encuentre:

(a) la moda muestral

(b) la mediana muestral

(c) la media muestral

Solución

(a) Puesto que el valor 4 aparece con la mayor frecuencia, la moda muestral es 4.

(b) Puesto que existen 30 valores de datos, la mediana muestral coincide con la media entre el 15º y el 16º valor menor. Puesto que el 15º valor menor es 3 y el 16º valor menor es 4, la mediana muestral es 3,5.

(c) La media muestral es

$$\bar{x} = \frac{1 \cdot 6 + 2 \cdot 4 + 3 \cdot 5 + 4 \cdot 8 + 5 \cdot 3 + 6 \cdot 4}{30} = \frac{100}{30} \approx 3{,}333 \quad \blacksquare$$

Problemas

1. Relacione cada sentencia de la izquierda con el conjunto de datos correcto entre los que figuran a la derecha.

1. La moda muestral es 9	A: 5, 7, 8, 10, 13, 14
2. La media muestral es 9	B: 1, 2, 5, 9, 9, 15
3. La mediana muestral es 9	C: 1, 2, 9, 12, 12, 18

2. Con los datos del ejemplo 2.2, encuentre la moda muestral de las puntuaciones ganadoras del Torneo de Maestros de Golf.

3. Con los datos de los primeros 100 estudiantes del Apéndice A, encuentre la moda muestral para:

   (a) los pesos

   (b) las presiones sanguíneas

   (c) los niveles de colesterol

4. Suponga que usted quiere descubrir el salario del vicepresidente de un banco, al que acaba de conocer. Si pretende tener la mayor posibilidad de acertar a menos de 1000 dólares, ¿le gustaría conocer la media muestral, la mediana muestral o la moda muestral de los salarios de los vicepresidentes de bancos?

5. Construya un conjunto de datos para el que la media muestral sea 10, la mediana muestral sea 8 y la moda muestral sea 6.

6. Si la moda muestral de un conjunto de datos x_i, $i = 1, \ldots, n$, es igual a 10, ¿cuál será la moda muestral de los datos $y_i = 2x_i + 5$, $i = 1, \ldots, n$?

7. Varios corredores amateurs utilizan una pista de atletismo de un cuarto de milla de longitud. En una muestra de 17 corredores, 1 corrió 2 vueltas, 4 corrieron 4 vueltas, 5 corrieron 6 vueltas, 6 corrieron 8 vueltas y 1 corrió 12 vueltas.

   (a) ¿Cuál es la moda muestral del número de vueltas que han hecho estos corredores?

   (b) ¿Cuál es la moda muestral de las distancias en millas recorridas por los corredores?

3.5 Varianza muestral y desviación típica muestral

Aunque hasta ahora se han introducido estadísticos que miden la tendencia central de un conjunto de datos, todavía no se han considerado aquellos que miden su dispersión o variabilidad. Por ejemplo, pese a que los siguientes conjuntos de datos A y B tienen las mismas media y mediana muestrales, claramente existe una mayor dispersión en los valores de B que en los de A.

$$A\text{: } 1, 2, 5, 6, 6 \qquad B\text{: } -40, 0, 5, 20, 35$$

Una forma de medir la variabilidad de un conjunto de datos consiste en considerar las desviaciones de los valores de datos a un valor central. El valor central que se utiliza más frecuentemente para este propósito es la media muestral. Si los valores de datos son $x_1, \ldots, x_n$ y la media muestral es $\bar{x} = \sum_{i=1}^{n} x_i/n$, la desviación a la media del valor x_i es , $i = 1, \ldots, n$.

Se podría suponer que una medida natural de la variabilidad de un conjunto de datos es la media de las desviaciones a la media. Sin embargo, como se ha visto en la sección 3.2, $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$. Esto es, la suma de las desviaciones a la media muestral es siempre igual a 0 y, por consiguiente, la media de las desviaciones a la media muestral también será igual a 0. Ahora bien, tras una reflexión adicional, se verá claro que no se desea permitir que las desviaciones positivas y negativas se compensen. Por el contrario, se deberían considerar las desviaciones a la media sin tener en cuenta sus signos. Esto se puede conseguir si se consideran los valores absolutos de las desviaciones o, algo más útil, si se consideran sus cuadrados.

La varianza muestral se define como la "media" de los cuadrados de las desviaciones a la media muestral. Sin embargo, por cuestiones técnicas (que se verán claras en el capítulo 8), esta "media" divide la suma de las n desviaciones al cuadrado por $n - 1$, en lugar de dividirla por n, como es habitual.

Definición

La *varianza muestral*, denotada por s^2, de los datos $x_i, \ldots, x_n$ con media $\bar{x} = \left(\sum_{i=1}^{n} x_i\right)/n$ se define como

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$$

Ejemplo 3.15 Encuentre la varianza muestral del conjunto de datos A.

Solución Se determinará como sigue:

x_i	1	2	5	6	6
$\bar{x}$	4	4	4	4	4
$x_i - \bar{x}$	−3	−2	1	2	2
$(x_i - \bar{x})^2$	9	4	1	4	4

De donde, para el conjunto de datos A,

$$s^2 = \frac{9 + 4 + 1 + 4 + 4}{4} = 5{,}5 \quad \blacksquare$$

Ejemplo 3.16 Encuentre la varianza muestral del conjunto de datos B.

Solución La media muestral del conjunto de datos B es también $\bar{x} = 4$. Por consiguiente, para este conjunto de datos, se tendrá

x_i	−40	0	5	20	35
$x_i - \bar{x}$	−44	−4	1	16	31
$(x_i - \bar{x})^2$	1936	16	1	256	961

Así pues,

$$s^2 = \frac{3170}{4} = 792{,}5 \quad \blacksquare$$

La siguiente identidad algebraica resulta útil cuando se desea calcular a mano la varianza muestral:

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \tag{3.2}$$

Ejemplo 3.17 Compruebe que, para el conjunto de datos A, se verifica la identidad (3.2).

Solución Puesto que $n = 5$ y $\bar{x} = 4$,

$$\sum_{i=1}^{5} x_i^2 - n\bar{x}^2 = 1 + 4 + 25 + 36 + 36 - 5(16) = 102 - 80 = 22$$

Del ejemplo 3.15,

$$\sum_{i=1}^{5} (x_i - \bar{x})^2 = 9 + 4 + 1 + 4 + 4 = 22$$

y, por consiguiente, la identidad queda comprobada. $\blacksquare$

Supongamos que se suma una constante c a cada uno de los valores de datos $x_1, \ldots, x_n$ para así obtener un nuevo conjunto de datos $y_1, \ldots, y_n$, donde

$$y_i = x_i + c$$

Para ver cómo afecta esto a la varianza muestral, recuerde, de la sección 3.2, que

$$\bar{y} = \bar{x} + c$$

y, por tanto,

$$y_i - \bar{y} = x_i + c - (\bar{x} + c) = x_i - \bar{x}$$

Es decir, las desviaciones y son iguales a las desviaciones x; en consecuencia, sus respectivas sumas de cuadrados son iguales. Así pues, se ha demostrado el resultado siguiente:

La varianza muestral se mantiene constante cuando se suma una constante a cada valor de dato

Se puede utilizar el resultado anterior, junto con la identidad (3.2), para reducir enormemente el tiempo de cálculo de la varianza muestral.

Ejemplo 3.18 Los siguientes datos muestran el número de policías asesinados en actos de servicio en Estados Unidos a lo largo de 10 años.

164, 165, 157, 164, 152, 147, 148, 131, 147, 155

Encuentre la varianza muestral del número de policías asesinados en esos años.

Solución En vez de trabajar directamente con los datos dados, restemos el valor 150 de cada uno de ellos. (Esto es, sumemos $c = -150$ a cada valor de dato.) Así se obtiene el conjunto de datos nuevo:

14, 15, 7, 14, 2, −3, −2, −19, −3, 5

Su media muestral es

$$\bar{y} = \frac{14 + 15 + 7 + 14 + 2 - 3 - 2 - 19 - 3 + 5}{10} = 3{,}0$$

La suma de los cuadrados de los datos nuevos es

$$\sum_{i=1}^{10} y_i^2 = 14^2 + 15^2 + 7^2 + 14^2 + 2^2 + 3^2 + 2^2 + 19^2 + 3^2 + 5^2 = 1078$$

Así pues, si se usa la identidad (3.2), se llega a que

$$\sum_{i=1}^{10} (y_i - \bar{y})^2 = 1078 - 10(9) = 988$$

Por lo tanto, la varianza muestral de los datos nuevos, que coincide con la de los datos originales, es

$$s^2 = \frac{988}{9} \approx 109{,}78 \quad \blacksquare$$

La raíz cuadrada positiva de la varianza muestral se denomina *desviación típica muestral.*

Definición

Al valor s, definido por

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}$$

se le denomina *desviación típica muestral.*

La desviación típica muestral se mide en las mismas unidades que los datos originales. Es decir, si por ejemplo los datos originales están dados en pies de longitud, la varianza muestral vendrá expresada en pies al cuadrado; mientras que la desviación típica muestral vendrá dada en pies.

Si cada valor de dato x_i, $i = 1, \ldots, n$, se multiplica por una constante c para obtener el nuevo conjunto de datos

$$y_i = cx_i \quad i = 1, \ldots, n$$

la varianza muestral de los datos y coincide con la varianza muestral de los datos x multiplicada por c^2. Esto es,

$$s_y^2 = c^2 s_x^2$$

donde s_y^2 y s_x^2 son las varianzas muestrales de los datos nuevos y de los datos originales, respectivamente. Si se extrae la raíz cuadrada de los dos miembros de la igualdad anterior se obtiene que la desviación típica de los datos y es igual a la desviación típica de los datos x multiplicada por el valor absoluto de c; es decir,

$$s_y = |c| s_x$$

Otro indicador de la variabilidad de un conjunto de datos es el *rango intercuartílico*, que es igual a la diferencia entre el tercer y el primer cuartil. Esto es, hablando grosso modo, el rango intercuartílico es la longitud del intervalo que contiene la mitad central de los datos.

Tabla 3.2 Distintos percentiles del Test de Analogías de Miller para cinco tipos de estudiantes

Percentil de orden (en %)	Ciencias Físicas	Medicina	Ciencias Sociales	Lengua y Literatura	Derecho
99	93	92	90	87	84
90	88	78	82	80	73
75	80	71	74	73	60
50	68	57	61	59	49
25	55	45	49	43	37

Ejemplo 3.19 El Test de Analogías de Miller es un test estandarizado al que se someten distintos alumnos que intentan acceder a determinados estudios universitarios y profesionales. La tabla 3.2 muestra algunos de los percentiles de las puntuaciones obtenidas por algunos de los estudiantes que se han presentado al test, clasificados por el tipo de estudios que pretenden seguir. Por ejemplo, la tabla 3.2 indica que 68 es la puntuación mediana de los estudiantes de Ciencias Físicas, mientras que la mediana de los de Derecho es 49.

Determine los rangos intercuartílicos para los estudiantes de los cinco tipos de estudios especificados.

Solución Puesto que el rango intercuartílico es la diferencia entre los percentiles de órdenes 75 y 25%, se tendrá que su valor será

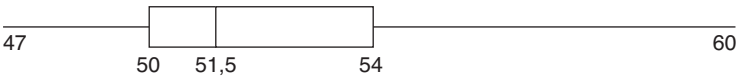
$80 - 55 = 25$ para las puntuaciones de los estudiantes de Ciencias Físicas
$71 - 45 = 26$ para las puntuaciones de los estudiantes de Medicina
$74 - 49 = 25$ para las puntuaciones de los estudiantes de Ciencias Sociales
$73 - 43 = 30$ para las puntuaciones de los estudiantes de Lengua y Literatura
$60 - 37 = 23$ para las puntuaciones de los estudiantes de Derecho ■

Los *diagramas de caja* se utilizan habitualmente para representar algunos de los estadísticos sintéticos de un conjunto de datos. En el eje x se dibuja un segmento entre los valores menor y mayor de los datos; superpuesta a este segmento, se coloca una "caja" que comienza en el primer cuartil y termina en el tercer cuartil, dentro de la cual se indica el valor del segundo cuartil mediante una línea vertical. Por ejemplo, en la siguiente tabla de frecuencias se muestran los salarios iniciales de una muestra de 42 graduados en Arte.

Salario inicial	Frecuencia
47	4
48	1
49	3
50	5
51	8

Salario inicial	Frecuencia
52	10
53	0
54	5
56	2
57	3
60	1

Los salarios oscilan entre los valores menor y mayor que coinciden con 47 y 60, respectivamente. El valor del primer cuartil (igual al 11º menor salario de la lista) es 50; el valor del segundo cuartil (igual a la media entre el 21º y 22º menores salarios) es 51,5; y el valor del cuartil tercero (que coincide con el 32º menor salario de la lista) es 54. El diagrama de caja para este conjunto de datos es el siguiente:



Un diagrama de caja.

Problemas

1. En la tabla siguiente se muestran los consumos per cápita de leche, en los años comprendidos entre 1983 y 1987, en Estados Unidos. Los datos provienen del Departamento de Agricultura de Estados Unidos, *Consumo de alimentos, precios y gastos*, anuario.

Año	Consumo (en galones per cápita)
1983	26,3
1984	26,2
1985	26,4
1986	26,3
1987	25,9

Encuentre la media muestral y la varianza muestral para estos datos.

2. Considere los dos conjuntos de datos siguientes:

A: 66, 68, 71, 72, 72, 75 *B*: 2, 5, 9, 10, 10, 16

(a) ¿Cuál parece tener mayor varianza muestral?

(b) Determine la varianza muestral del conjunto de datos A.

(c) Determine la varianza muestral del conjunto de datos B.

3. Los torneos de Maestros de Golf y Abierto de Estados Unidos son dos de los más prestigiosos torneos de golf de Estados Unidos. El torneo de Maestros se juega siempre en el campo de golf de Augusta, mientras que el Abierto se juega en diferentes campos cada año. Por ello, es probable que la varianza muestral de las puntuaciones ganadoras del Abierto de Estados Unidos sea mayor que la de las puntuaciones del torneo de Maestros. Para comprobar si es así se han recopilado las puntuaciones ganadoras de ambos torneos durante los años comprendidos entre 1981 y 1990.

	Puntuación ganadora									
Torneo	**1981**	**1982**	**1983**	**1984**	**1985**	**1986**	**1987**	**1988**	**1989**	**1990**
Abierto de EU	273	282	280	276	279	279	277	278	278	280
Maestros de Golf	280	284	280	277	282	279	285	281	283	278

(a) Calcule la varianza muestral de las puntuaciones ganadoras del Torneo Abierto de Estados Unidos.

(b) Calcule la varianza muestral de las puntuaciones ganadoras del Torneo de Maestros de Golf.

La tabla siguiente muestra el número de médicos y dentistas que había Japón en los años pares comprendidos entre 1984 y 2000. Los problemas 4 y 5 se refieren a esta tabla.

Número de médicos y dentistas (1984-2000)

	Médicos	**Dentistas**
1984	173 452	61 283
1986	183 129	64 904
1988	193 682	68 692
1990	203 797	72 087
1992	211 498	75 628
1994	220 853	79 091
1996	230 297	83 403
1998	236 933	85 669
2000	243 201	88 410

4. Determine la varianza muestral del número de médicos en los años citados.

5. Determine la varianza muestral del número de dentistas en dichos años.

6. Un individuo que necesitaba asegurar su coche preguntó cuáles eran las cuotas para idénticas coberturas en 10 compañías de seguros. Obtuvo los siguientes valores (correspondientes a las cuotas anuales, en dólares).

720, 880, 630, 590, 1140, 908, 677, 720, 1260, 800

Encuentre:

(a) la media muestral

(b) la mediana muestral

(c) la desviación típica muestral

La siguiente tabla muestra la población de 2003 en cada uno de los 50 Estados y en el Distrito de Columbia de Estados Unidos. Los problemas 7, 8 y 9 se refieren a esta tabla.

Población residente, 1 de julio de 2003

Estado	**Número**	**Rango de orden**
Estados Unidos	290 809 777	(X)
Alabama	4 500 752	23
Alaska	648 818	47
Arizona	5 580 811	18
Arkansas	2 725 714	32
California	35 484 453	1
Colorado	4 550 688	22
Connecticut	3 483 372	29
Delaware	817 491	45
District of Columbia	563 384	(X)
Florida	17 019 068	4
Georgia	8 684 715	9
Hawaii	1 257 608	42
Idaho	1 366 332	39
Illinois	12 653 544	5
Indiana	6 195 643	14
Iowa	2 944 062	30
Kansas	2 723 507	33
Kentucky	4 117 827	26
Louisiana	4 496 334	24
Maine	1 305 728	40
Maryland	5 508 909	19
Massachusetts	6 433 422	13
Michigan	10 079 985	8
Minnesota	5 059 375	21
Mississippi	2 881 281	31
Missouri	5 704 484	17
Montana	917 621	44
Nebraska	1 739 291	38
Nevada	2 241 154	35
New Hampshire	1 287 687	41

(*Continúa*)

Población residente, 1 de julio de 2003

Estado	Número	Rango de orden
New Jersey	8 638 396	10
New Mexico	1 874 614	36
New York	19 190 115	3
North Carolina	8 407 248	11
North Dakota	633 837	48
Ohio	11 435 798	7
Oklahoma	3 511 532	28
Oregon	3 559 596	27
Pennsylvania	12 365 455	6
Rhode Island	1 076 164	43
South Carolina	4 147 152	25
South Dakota	764 309	46
Tennessee	5 841 748	16
Texas	22 118 509	2
Utah	2 351 467	34
Vermont	619 107	49
Virginia	7 386 330	12
Washington	6 131 445	15
West Virginia	1 810 354	37
Wisconsin	5 472 299	20
Wyoming	501 242	50

Observación: Cuando varios Estados comparten el mismo rango de orden, el siguiente rango se omite. Debido al redondeo de los datos, varios Estados pueden tener valores idénticos pero rangos distintos.

7. Encuentre la varianza muestral de las poblaciones de los primeros 17 Estados.
8. Encuentre la varianza muestral de las poblaciones de los siguientes 17 Estados.
9. Encuentre la varianza muestral de las poblaciones de los últimos 17 Estados.
10. Si s^2 es la varianza muestral de los datos x_i, $i = 1, \ldots, n$, ¿cuál es la varianza muestral de los datos $ax_i + b$, $i = 1, \ldots, n$, donde a y b son constantes dadas?
11. Calcule la varianza muestral y la desviación típica muestral de los siguientes conjuntos de datos:
    (a) 1, 2, 3, 4, 5
    (b) 6, 7, 8, 9, 10
    (c) 11, 12, 13, 14, 15
    (d) 2, 4, 6, 8, 10
    (e) 10, 20, 30, 40, 50
12. En el lado estadounidense de la frontera con Canadá las temperaturas se miden en grados Fahrenheit, mientras que en el lado canadiense se miden en grados Celsius (o cen-

tígrados). Supongamos que la temperatura media diaria registrada durante el mes de enero en el lado de Estados Unidos fue de 40 ºF y que la varianza muestral fue de 12.

Utilice la fórmula siguiente, que permite transformar temperaturas Fahrenheit a Celsius

$$C = \frac{5}{9}(F - 32)$$

para encontrar

(a) la media muestral registrada por los canadienses

(b) la varianza muestral registrada por los canadienses

13. Calcule la media muestral y la varianza muestral de las presiones sistólicas sanguíneas de los primeros 50 estudiantes del conjunto de datos del Apéndice A. Haga lo mismo con los últimos 50 estudiantes del citado conjunto de datos. Compare las respuestas. Comente los resultados de esta comparación. ¿Resultan sorprendentes?

14. Si s es la desviación típica muestral de los datos x_i, $i = 1, \ldots, n$, ¿cuál es la desviación típica muestral de $ax_i + b$, $i = 1, \ldots, n$? En este problema, a y b son constantes dadas.

15. La siguiente tabla muestra el número de motos vendidas en Estados Unidos durante 8 años distintos. Utilícela para calcular la desviación típica muestral de las ventas de motos en los años citados.

Año	1975	1980	1983	1984	1985	1986	1987	1988
Ventas de motos (en miles)	940	1070	1185	1305	1260	1045	935	710

Fuente: Consejo de la Industria de Motocicletas.

16. Encuentre la desviación típica muestral del conjunto de datos reflejado en la siguiente tabla de frecuencias:

Valor	**Frecuencia**	**Valor**	**Frecuencia**
3	1	5	3
4	2	6	2

17. Los datos siguientes representan la acidez de 40 precipitaciones de lluvia sucesivas en el estado de Minnesota. La acidez se mide en una escala de pH que varía de 1 (muy ácida) a 7 (neutra).

3,71, 4,23, 4,16, 2,98, 3,23, 4,67, 3,99, 5,04, 4,55, 3,24, 2,80, 3,44, 3,27, 2,66, 2,95, 4,70, 5,12, 3,77, 3,12, 2,38, 4,57, 3,88, 2,97, 3,70, 2,53, 2,67, 4,12, 4,80, 3,55, 3,86, 2,51, 3,33, 3,85, 2,35, 3,12, 4,39, 5,09, 3,38, 2,73, 3,07

(a) Encuentre la desviación típica muestral.

(b) Obtenga el rango muestral.

(c) Encuentre el rango intercuartílico.

18. Considere los dos siguientes conjuntos de datos.

$$A\text{: } 4{,}5,\ 0,\ 5{,}1,\ 5{,}0,\ 10,\ 5{,}2 \qquad B\text{: } 0{,}4,\ 0{,}1,\ 9,\ 0,\ 10,\ 9{,}5$$

(a) Determine el rango de cada conjunto de datos.

(b) Calcule la desviación típica muestral de cada conjunto de datos.

(c) Determine el rango intercuartílico de cada conjunto de datos.

3.6 Conjuntos de datos normales y la regla empírica

En la práctica, la mayoría de los conjuntos de datos grandes que uno encuentra tienen histogramas similares en cuanto a la forma. Por lo general, esos histogramas son simétricos con respecto al punto de máxima frecuencia y decrecen a ambos lados de ese punto siguiendo una forma acampanada. Tales conjuntos de datos se dice que son *normales*, y sus histogramas se denominan *histogramas normales*.

Definición

Se dice que un conjunto de datos es *normal* si el histograma que lo describe tiene las propiedades siguientes:

1. La máxima altura se alcanza en el intervalo central.
2. Si nos movemos desde el intervalo central en cualquier dirección, la altura decrece de tal modo que el histograma completo tiene una forma acampanada.
3. El histograma es simétrico con respecto al intervalo central.

La figura 3.2 muestra el histograma de un conjunto de datos normal.

Si el histograma de un conjunto de datos se aproxima al de un histograma normal, se dice que el conjunto de datos es *aproximadamente normal*. Por ejemplo, el histograma que aparece en la figura 3.3 proviene de un conjunto de datos aproximadamente normal; mientras que los histogramas presentados de las figuras 3.4 y 3.5 no lo son (puesto que cada uno de ellos es manifiestamente asimétrico). Cualquier conjunto de datos que no sea simétrico con respecto a su mediana se dice que es *asimétrico*. Se dice que es *asimétrico por la derecha* si tiene una cola alargada por la derecha, y se dice que es *asimétrico por la izquierda* si la cola alargada se encuentra a la izquierda. Así pues, el conjunto de datos representado en la figura 3.4 es asimétrico por la izquierda, mientras que el representado en la figura 3.5 es asimétrico por la derecha.

Se desprende de la simetría de los histogramas normales que, si un conjunto de datos es aproximadamente normal, su media muestral y su mediana muestral son aproximadamente iguales.

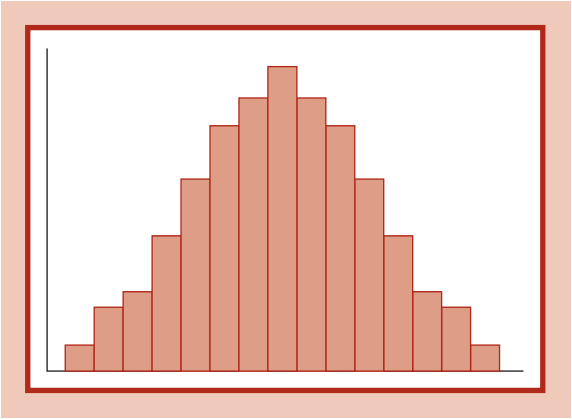


Figura 3.2 Histograma de un conjunto de datos normal.

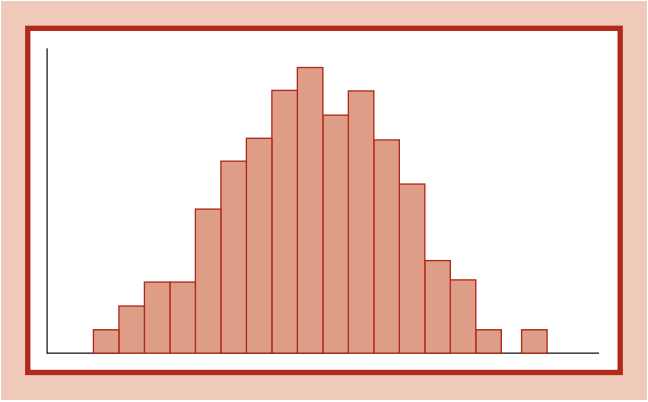


Figura 3.3 Histograma de un conjunto de datos aproximadamente normal.

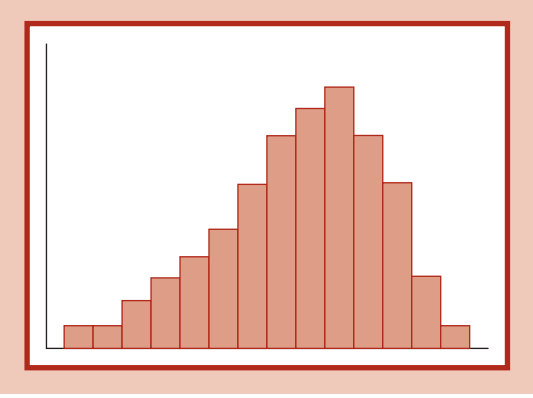


Figura 3.4 Histograma de un conjunto de datos asimétrico por la izquierda.

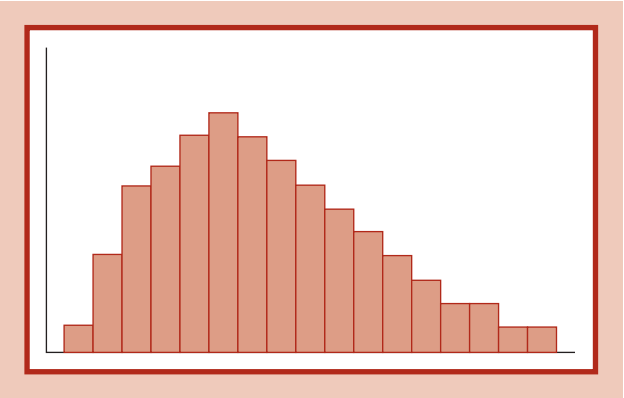


Figura 3.5 Histograma de un conjunto de datos asimétrico por la derecha.

Supongamos que $\bar{x}$ y s son, respectivamente, la media muestral y la desviación típica muestral de un conjunto de datos aproximadamente normal. La regla siguiente, conocida como *regla empírica*, especifica las proporciones aproximadas de datos observados que distan de la media muestral $\bar{x}$ en menos de s, $2s$ y $3s$.

Regla empírica

Si un conjunto de datos es aproximadamente normal con media muestral $\bar{x}$ y desviación típica muestral s, los siguientes puntos son ciertos:

1. Aproximadamente, un 68% de las observaciones caen dentro de

$$\bar{x} \pm s$$

2. Aproximadamente, un 95% de las observaciones caen dentro de

$$\bar{x} \pm 2s$$

3. Aproximadamente, un 99,7% de las observaciones caen dentro de

$$\bar{x} \pm 3s$$

Ejemplo 3.20 Las calificaciones obtenidas por 25 estudiantes en un examen de Historia aparecen representadas en el siguiente diagrama de tallos y hojas.

9	0, 0, 4
8	3, 4, 4, 6, 6, 9
7	0, 0, 3, 5, 5, 8, 9
6	2, 2, 4, 5, 7
5	0, 3, 5, 8

Si miramos esta figura de lado (o, equivalentemente, si giramos el libro) se puede ver que el histograma correspondiente es aproximadamente normal. Utilícela para evaluar la regla empírica.

Solución Si se hacen los cálculos pertinentes se obtiene que la media muestral y la desviación típica muestral son

$$\bar{x} = 73{,}68 \quad \text{y} \quad s = 12{,}80$$

La regla empírica establece que aproximadamente un 68% de los valores de datos se encuentran entre $\bar{x} - s = 60{,}88$ y $\bar{x} + s = 86{,}48$. Puesto que 17 observaciones caen realmente entre 60,88 y 86,48, el porcentaje real es del $100(17/25) = 68\%$. Del mismo modo, la regla empírica establece que aproximadamente un 95% de los datos se encuentran entre $\bar{x} - 2s = 48{,}08$ y $\bar{x} + 2s = 96{,}28$; mientras que realmente el 100% de los datos se encuentran dentro de este rango. ■

Un conjunto de datos que se ha obtenido muestreando una sobre población compuesta por varias subpoblaciones de tipos diferentes no es, por lo general, normal. Por el contrario, el histograma de tal conjunto de datos parece reflejar una combinación, o superposición, de histogramas normales y, en consecuencia, suele tener más de un pico, o chepa, local. Debido a que el histograma será más alto en esos picos locales que en otros valores próximos a ellos, esos picos son similares a las modas. Un conjunto de datos cuyo histograma tenga dos picos locales se dice que es *bimodal*. El conjunto de datos representado en la figura 3.6 es bimodal.

Puesto que un gráfico de tallos y hojas puede ser considerado como un histograma girado, aquél es útil para observar si un conjunto de datos es aproximadamente normal.

Ejemplo 3.21 El siguiente gráfico de tallos y hojas representa los pesos de 200 miembros de un club de salud.

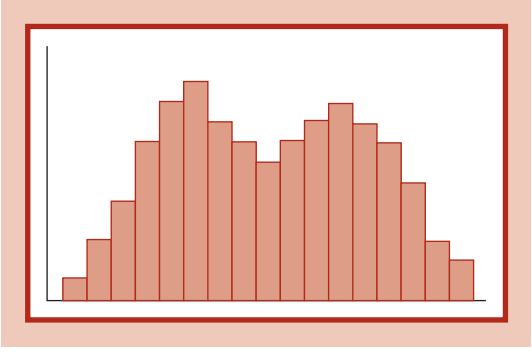


Figura 3.6 Histograma de un conjunto de datos bimodal.

24	9
23	
22	1
21	7
20	2, 2, 5, 5, 6, 9, 9, 9
19	0, 0, 0, 0, 0, 1, 1, 2, 4, 4, 5, 8
18	0, 1, 1, 2, 2, 2, 3, 4, 4, 4, 5, 5, 5, 6, 6, 6, 6, 7, 9, 9, 9
17	1, 1, 1, 2, 3, 3, 4, 4, 4, 5, 5, 6, 6, 6, 6, 7, 7, 7, 7, 9
16	0, 0, 1, 1, 1, 1, 2, 4, 5, 5, 6, 6, 8, 8, 8, 8
15	0, 1, 1, 1, 1, 1, 1, 5, 5, 5, 5, 6, 6, 6, 7, 7, 8, 9
14	0, 0, 0, 1, 2, 3, 4, 5, 6, 7, 7, 7, 8, 9, 9
13	0, 0, 0, 1, 1, 1, 2, 2, 2, 2, 3, 3, 4, 5, 5, 6, 6, 6, 6, 7, 7, 8, 8, 8, 9, 9, 9
12	1, 1, 1, 2, 2, 2, 3, 4, 4, 5, 5, 6, 6, 6, 6, 6, 6, 6, 6, 7, 7, 7, 7, 8, 8, 9, 9, 9
11	0, 1, 1, 2, 2, 2, 2, 3, 3, 4, 4, 5, 5, 6, 9, 9
10	0, 2, 3, 3, 3, 4, 4, 5, 7, 7, 8
9	0, 0, 9
8	6

Si se observa de lado, se ve que su histograma no parece aproximadamente normal. Sin embargo, es importante resaltar que estos datos consisten en los pesos de todos los miembros del club de salud, tanto mujeres como hombres. Puesto que estos dos grupos determinan dos poblaciones diferentes en cuanto a sus pesos, tiene sentido considerar separadamente los datos de cada sexo. Esto se hará a continuación.

Resulta que estos 200 valores de datos son los pesos de 97 mujeres y de 103 hombres. Si se separan los pesos de las mujeres de los pesos de los hombres, se obtienen los gráficos de tallos y hojas de las figuras 3.7 y 3.8.

Como se puede ver en estas figuras, parece que los datos separados por sexo son aproximadamente normales. Calculemos $\bar{x}_w$, s_w, $\bar{x}_m$ y s_m, las medias muestrales y las desviaciones típicas muestrales de las mujeres y los hombres, respectivamente.

16	0, 5
15	0, 1, 1, 1, 5
14	0, 0, 1, 2, 3, 4, 6, 7, 9
13	0, 0, 1, 1, 2, 2, 2, 2, 3, 4, 5, 5, 6, 6, 6, 6, 7, 8, 8, 8, 9, 9, 9
12	1, 1, 1, 2, 2, 2, 3, 4, 4, 5, 5, 6, 6, 6, 6, 6, 6, 6, 6, 7, 7, 7, 7, 8, 8, 9, 9
11	0, 0, 1, 1, 2, 2, 2, 2, 3, 3, 4, 4, 5, 5, 6, 9, 9
10	2, 3, 3, 3, 4, 4, 5, 7, 7, 8
9	0, 0, 9
8	6

Figura 3.7 Pesos de las 97 mujeres del club de salud.

24	9
23	
22	1
21	7
20	2, 2, 5, 5, 6, 9, 9, 9
19	0, 0, 0, 0, 0, 1, 1, 2, 4, 4, 5, 8
18	0, 1, 1, 2, 2, 2, 3, 4, 4, 4, 5, 5, 5, 6, 6, 6, 6, 7, 9, 9, 9
17	1, 1, 1, 2, 3, 3, 4, 4, 4, 5, 5, 6, 6, 6, 6, 7, 7, 7, 7, 9
16	0, 1, 1, 1, 1, 2, 4, 5, 6, 6, 8, 8, 8, 8
15	1, 1, 1, 5, 5, 5, 6, 6, 6, 7, 7, 8, 9
14	0, 5, 7, 7, 8, 9
13	0, 1, 2, 3, 7
12	9

Figura 3.8 Pesos de los 103 hombres del club de salud.

Los cálculos conducen a

$$\bar{x}_w = 125{,}70 \quad \bar{x}_m = 174{,}69$$
$$s_w = 15{,}58 \quad s_m = 21{,}23$$

Una comprobación adicional de la normalidad aproximada de los dos conjuntos de datos separados se obtiene si se observa la similitud entre los valores de la media muestral y de la mediana muestral de ambos casos. La mediana muestral de los pesos de las mujeres coincide con el 49º menor valor de dato, que es igual a 126; mientras que la mediana muestral de los datos de los hombres coincide con el 52º valor menor, que es igual a 174. Ambas medianas están próximas a sus correspondientes medias muestrales, cuyos valores son 125,7 y 174,69.

Dados los valores de la media muestral y de la desviación típica muestral, de la regla empírica se deduce que aproximadamente un 68% de las mujeres tendrán unos pesos comprendidos entre 110,1 y 141,3, y que aproximadamente un 95% de los hombres tendrán unos pesos comprendidos entre 132,2 y 217,2. De las figuras 3.7 y 3.8 se puede comprobar que los porcentajes reales son

$$100 \times \frac{68}{97} = 70{,}1 \quad \text{y} \quad 100 \times \frac{101}{103} = 98{,}1 \quad \blacksquare$$

Problemas

1. Los datos siguientes muestran el número de animales tratados diariamente en una clínica veterinaria a lo largo de un periodo de 24 días:

22, 17, 19, 31, 28, 29, 21, 33, 36, 24, 15, 28, 25, 28, 22, 27, 33, 19, 25, 28, 26, 20, 30, 32

(North Wind Picture Archives)

Adolphe Quetelet

Perspectiva histórica

Quetelet y el fraude descubierto mediante la curva normal

El estadístico y científico social belga Adolphe Quetelet fue un gran defensor de la hipótesis de que la mayor parte de los conjuntos de datos referidos a medidas humanas eran normales. En un estudio, midió el torso de 5738 soldados escoceses, representó gráficamente el conjunto de datos resultante en un histograma y concluyó que era normal.

En un posterior estudio, Quetelet utilizó la forma de los histogramas normales para descubrir la evidencia de un fraude relacionado con los reclutas del ejército francés. Estudió los datos relativos a las alturas de una extensa muestra de 100 000 reclutas. Representó gráficamente los datos en un histograma –con intervalos de clase de 1 pulgada– y encontró que, los datos parecían ser normales, con la excepción de tres intervalos de clase de alrededor de 62 pulgadas. En particular, existían menos valores en el intervalo comprendido entre 62 y 63 pulgadas; mientras que, en los intervalos de 60 a 61 y de 61 a 62 pulgadas, había ligeramente más de los que cabría esperar en un ajuste normal perfecto de los datos. Intentando averiguar por qué la curva normal no se ajustaba tan bien a los datos como él había supuesto que lo haría, Quetelet descubrió que 62 pulgadas era la altura mínima exigida a los soldados del ejército francés. Basándose en esto y en su idea sobre la muy extensa aplicabilidad de los datos normales, Quetelet llegó a la conclusión de que algunos reclutas, cuyas alturas eran ligeramente superiores a 62 pulgadas, "doblaban sus rodillas" para parecer más bajos y evitar, así, su reclutamiento.

Durante los siguientes 50 años posteriores a Quetelet, grosso modo entre 1840 y 1890, estuvo ampliamente extendida la idea de que la mayoría de los conjuntos de datos procedentes de poblaciones homogéneas (es decir, datos que claramente no provinieran de una mixtura de poblaciones diferentes) deberían ser normales, siempre que los tamaños muestrales fueran lo suficientemente grandes. Aunque los estadísticos actuales en cierto modo se muestran escépticos respecto a esa idea, es bastante corriente el que un conjunto de datos provenga de una población normal. Este fenómeno, que generalmente ocurre en los conjuntos de datos que surgen en las ciencias biológicas y físicas, se puede explicar en parte mediante un resultado matemático conocido como *teorema central del límite*. En realidad, el teorema central del límite (que se estudia en el capítulo 7) explicará por sí mismo por qué muchos de los conjuntos de datos que aparecen en las ciencias físicas son aproximadamente normales. Para explicar por qué, a menudo, los datos biométricos (es decir, datos generados en estudios de Biología) parecen ser normales, se utilizará una observación empírica debida a Francis Galton, conocida como regresión a la media, y que en la actualidad tiene una clara justificación científica. La *regresión a la media*, junto con el teorema central del límite y el paso de un gran número de generaciones, puede explicar por qué los conjuntos de datos biométricos son, habitualmente, normales. Esta explicación se presenta en el capítulo 12.

(a) Represente gráficamente estos datos en un histograma.

(b) Encuentre la media muestral.

(c) Encuentre la mediana muestral.

(d) ¿Son estos datos son aproximadamente normales?

2. Los datos siguientes reflejan las tasas de accidentalidad laboral por 100 000 horas trabajadas, para una muestra de empresas de semiconductores.

   1,4, 2,4, 3,7, 3,1, 2,0, 1,9, 2,5, 2,8, 2,2, 1,7, 3,1, 4,0, 2,2, 1,8, 2,6, 3,6, 2,9, 3,3, 2,0, 2,4

   (a) Represente gráficamente estos datos en un histograma.

   (b) ¿Este conjunto de datos es, a grandes rasgos, simétrico?

   (c) Si la respuesta a (b) es no, ¿es asimétrico por la izquierda o por la derecha?

   (d) Si la respuesta a (b) es sí, ¿es aproximadamente normal?

La tabla siguiente muestra las tasas de mortalidad infantil por 1000 nacimientos vivos en los 50 Estados y en el Distrito de Columbia de Estados Unidos. Los problemas 3 y 4 se refieren a esta tabla.

Tasa de mortalidad infantil, 2001

Estado	**Tasa**	**Rango de orden**
Estados Unidos	6,8	(X)
Alabama	9,4	4
Alaska	8,1	11
Arizona	6,9	26
Arkansas	8,3	10
California	5,4	45
Colorado	5,8	39
Connecticut	6,1	34
Delaware	10,7	1
District of Columbia	10,6	(X)
Florida	7,3	21
Georgia	8,6	8
Hawaii	6,2	32
Idaho	6,2	32
Illinois	7,7	14
Indiana	7,5	17
Iowa	5,6	43
Kansas	7,4	18
Kentucky	5,9	36
Louisiana	9,8	3
Maine	6,1	34
Maryland	8,1	11
Massachusetts	5,0	48
Michigan	8,0	13
Minnesota	5,3	47
Mississippi	10,5	2
Missouri	7,4	18
Montana	6,7	29

(*Continúa*)

Tasa de mortalidad infantil, 2001 (*Continuación*)

Estado	Tasa	Rango de orden
Nebraska	6,8	27
Nevada	5,7	42
New Hampshire	3,8	50
New Jersey	6,5	30
New Mexico	6,4	31
New York	5,8	39
North Carolina	8,5	9
North Dakota	8,8	6
Ohio	7,7	14
Oklahoma	7,3	21
Oregon	5,4	45
Pennsylvania	7,2	23
Rhode Island	6,8	27
South Carolina	8,9	5
South Dakota	7,4	18
Tennessee	8,7	7
Texas	5,9	36
Utah	4,8	49
Vermont	5,5	44
Virginia	7,6	16
Washington	5,8	39
West Virginia	7,2	23
Wisconsin	7,1	25
Wyoming	5,9	36

Observación: Representa las muertes de niños con una edad de menos de 1 año por cada 1000 nacimientos vivos, según el lugar de residencia. Excluye las muertes fetales. Cuando varios Estados comparten el mismo rango de orden, el siguiente rango se omite. Debido al redondeo de datos, varios Estados pueden tener valores idénticos pero rangos diferentes.

3. Para los datos sobre la mortalidad infantil.
   (a) Calcule la media muestral.
   (b) Calcule la mediana muestral.
4. Para los datos sobre la mortalidad infantil.
   (a) Represente estos datos mediante un gráfico de tallos y hojas.
   (b) ¿El conjunto de datos es aproximadamente normal?
5. Los siguientes datos son una muestra de precios de venta de casas en una comunidad de clase media de California. Los datos están dados en miles de dólares.

166, 82, 175, 181, 169, 177, 180, 185, 159, 164, 170, 149, 188,
173, 170, 164, 158, 177, 173, 175, 190, 172

(a) Encuentre la media muestral.

(b) Encuentre la mediana muestral.

(c) Represente gráficamente los datos en un histograma.

(d) ¿El conjunto de datos es aproximadamente normal?

6. Los datos siguientes muestran la edad que tenían en el momento de su proclamación los 43 presidentes de Estados Unidos.

Presidente	Edad de proclamación
1. Washington	57
2. J. Adams	61
3. Jefferson	57
4. Madison	57
5. Monroe	58
6. J. Q. Adams	57
7. Jackson	61
8. Van Buren	54
9. W. Harrison	68
10. Tyler	51
11. Polk	49
12. Taylor	64
13. Fillmore	50
14. Pierce	48
15. Buchanan	65
16. Lincoln	52
17. A. Johnson	56
18. Grant	46
19. Hayes	54
20. Garfield	49
21. Arthur	50
22. Cleveland	47
23. B. Harrison	55
24. Cleveland	55
25. McKinley	54
26. T. Roosevelt	42
27. Taft	51
28. Wilson	56
29. Harding	55
30. Coolidge	51
31. Hoover	54

(*Continúa*)

Presidente	Edad de proclamación
32. F. Roosevelt	51
33. Truman	60
34. Eisenhower	62
35. Kennedy	43
36. L. Johnson	55
37. Nixon	56
38. Ford	61
39. Carter	52
40. Reagan	69
41. Bush, Sr.	64
42. Clinton	46
43. Bush, Jr.	54

(a) Encuentre la media muestral y la desviación típica muestral de este conjunto de datos.

(b) Dibuje un histograma de los datos dados.

(c) ¿Los datos parecen aproximadamente normales?

(d) Si la respuesta a (c) es sí, obtenga un intervalo para el que se pueda esperar que contiene el 95% de los datos observados.

(e) ¿Qué porcentaje de datos cae realmente dentro del intervalo que se ha obtenido en el apartado (d)?

7. Para los datos sobre los pesos de mujeres del club de salud presentados en la figura 3.7 se calcularon la media muestral y la desviación típica muestral, que resultaron ser 125,70 y 15,58, respectivamente. Basándose en la forma mostrada en la figura 3.7 y en los valores anteriores, calcule la proporción aproximada de las mujeres con unos pesos comprendidos entre 94,54 y 156,86 libras. ¿Cuál es la proporción real?

8. Con una muestra de 36 varones enfermos del corazón se obtuvieron los siguientes datos relativos a las edades en las que sufrieron el primer ataque de corazón.

7	1, 2, 4, 5
6	0, 1, 2, 2, 3, 4, 5, 7
5	0, 1, 2, 3, 3, 4, 4, 4, 5, 6, 7, 8, 9
4	1, 2, 2, 3, 4, 5, 7, 8, 9
3	7, 9

(a) Determine $\bar{x}$ y s.

(b) A partir de la forma del gráfico de tallos y hojas, ¿qué porcentaje de valores de datos cabría esperar que estuvieran comprendidos entre $\bar{x} - s$ y $\bar{x} + s$? ¿Y entre $\bar{x} - 2s$ y $\bar{x} + 2s$?

(c) Encuentre los porcentajes reales para los intervalos dados en (b).

9. Si un histograma es asimétrico por la derecha, ¿qué estadístico será mayor: la media muestral o la mediana muestral? (*Sugerencia*: Si no está seguro, construya un conjunto de datos que sea asimétrico por la derecha y calcule después su media muestral y su mediana muestral.)

10. Los datos siguientes muestran las edades de 36 víctimas por crímenes violentos en una gran ciudad del este:

    25, 16, 14, 22, 17, 20, 15, 18, 33, 52, 70, 38, 18, 13, 22, 27, 19, 23,
    33, 15, 13, 62, 21, 57, 66, 16, 24, 22, 31, 17, 20, 14, 26, 30, 18, 25

    (a) Determine la media muestral.

    (b) Encuentre la mediana muestral.

    (c) Determine la desviación típica muestral.

    (d) ¿Este conjunto de datos parece aproximadamente normal?

    (e) ¿Qué proporción de datos dista de la media muestral menos de una vez la desviación típica muestral?

    (f) Compare la contestación dada en (e) con la aproximación derivada de la regla empírica.

La tabla siguiente lista las rentas per cápita, en 2002, para los 50 Estados de Estados Unidos. Los problemas de 11 a 13 se refieren a ella.

Rentas personales *per cápita* en dólares constantes (de 1996), año 2002

Estado	**Renta**	**Rango de orden**
Estados Unidos	27 857	(X)
Alabama	22 624	43
Alaska	28 947	14
Arizona	23 573	38
Arkansas	21 169	49
California	29 707	10
Colorado	29 959	9
Connecticut	38 450	1
Delaware	29 512	12
Florida	26 646	23
Georgia	25 949	28
Hawaii	27 011	20
Idaho	22 560	44
Illinois	30 075	8
Indiana	25 425	32

(*Continúa*)

Rentas personales per cápita en dólares constantes (de 1996), año 2002 (*Continuación*)

Estado	Renta	Rango de orden
Iowa	25 461	31
Kansas	26 237	26
Kentucky	23 030	39
Louisiana	22 910	41
Maine	24 979	33
Maryland	32 680	4
Massachusetts	35 333	3
Michigan	27 276	18
Minnesota	30 675	7
Mississippi	20 142	50
Missouri	26 052	27
Montana	22 526	45
Nebraska	26 804	22
Nevada	27 172	19
New Hampshire	30 912	6
New Jersey	35 521	2
New Mexico	21 555	47
New York	32 451	5
North Carolina	24 949	34
North Dakota	24 293	36
Ohio	26 474	25
Oklahoma	23 026	40
Oregon	25 867	29
Pennsylvania	28 565	15
Rhode Island	28 198	16
South Carolina	22 868	42
South Dakota	24 214	37
Tennessee	24 913	35
Texas	25 705	30
Utah	21 883	46
Vermont	26 620	24
Virginia	29 641	11
Washington	29 420	13
West Virginia	21 327	48
Wisconsin	26 941	21
Wyoming	27 530	17

Observación: Cuando varios Estados comparten el mismo rango de orden, el siguiente rango se omite. Debido al redondeo de datos, varios Estados pueden tener valores idénticos pero rangos diferentes.

11. Con los datos de los 25 primeros Estados:
    (a) Represente gráficamente los datos en un histograma.
    (b) Calcule la media muestral.
    (c) Calcule la mediana muestral.
    (d) Calcule la varianza muestral.
    (e) ¿Los datos son aproximadamente normales?
    (f) Utilice la regla empírica para obtener un intervalo que contenga aproximadamente el 68% de las observaciones.
    (g) Use la regla empírica para obtener un intervalo que contenga aproximadamente el 95% de las observaciones.
    (h) Determine la proporción real de observaciones que caen dentro del intervalo especificado en (f).
    (i) Determine la proporción real de observaciones que caen dentro del intervalo especificado en (g).
12. Repita el problema 11 utilizando, en esta ocasión, los datos de los 25 últimos Estados.
13. Repita el problema 11 utilizando ahora todos los datos de la tabla.

3.7 Coeficiente de correlación muestral

Consideremos el conjunto de datos apareados $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. En esta sección se presentará un estadístico, llamado *coeficiente de correlación muestral*, que mide el grado en el que valores grandes de x aparecen junto a valores grandes de y, y valores pequeños de x aparecen junto a valores pequeños de y.

Los datos de la tabla 3.3 reflejan los consumos medios de cigarrillos (variable x) y el número de radicales libres (variable y), medidos en las unidades adecuadas, que se han

Tabla 3.3 Consumo de cigarrillos y radicales libres

Persona	Número de cigarrillos consumidos	Radicales libres
1	18	202
2	32	644
3	25	411
4	60	755
5	12	144
6	25	302
7	50	512
8	15	223
9	22	183
10	30	375

encontrado en los pulmones de 10 fumadores. (Un radical libre es un solo átomo de oxígeno. Se cree que es potencialmente dañino porque es altamente reactivo y porque tiene una fuerte tendencia a combinarse con otros átomos dentro del cuerpo.) La figura 3.9 muestra un diagrama de dispersión de estos datos.

Si se examina la figura 3.9 se ve que cuando el consumo de cigarrillos es alto, el número de radicales libres tiende a ser igualmente alto, y que cuando el consumo de cigarrillos es bajo, el número de radicales libres también tiende a ser bajo. Cuando ocurre así, se dice que existe una *correlación positiva* entre las dos variables.

También estaremos interesados en determinar qué tipo de la relación existe entre dos variables cuando, en una de ellas, los valores altos tienden a estar asociados con los valores bajos en la otra. Por ejemplo, los datos de la tabla 3.4 representan los años de escolarización (variable x) y el pulso en situación de descanso, medido en latidos por minuto (variable y) para 10 individuos. En la figura 3.10 se incluye un diagrama de dispersión para estos datos. Se observa que los valores altos en el número de años de escolarización tienden a estar asociados con los valores bajos en el número de pulsaciones, y que los valores bajos en los años de escolarización tienden a estar asociados con los valores altos en el número de pulsaciones. Éste es un ejemplo de *correlación negativa*.

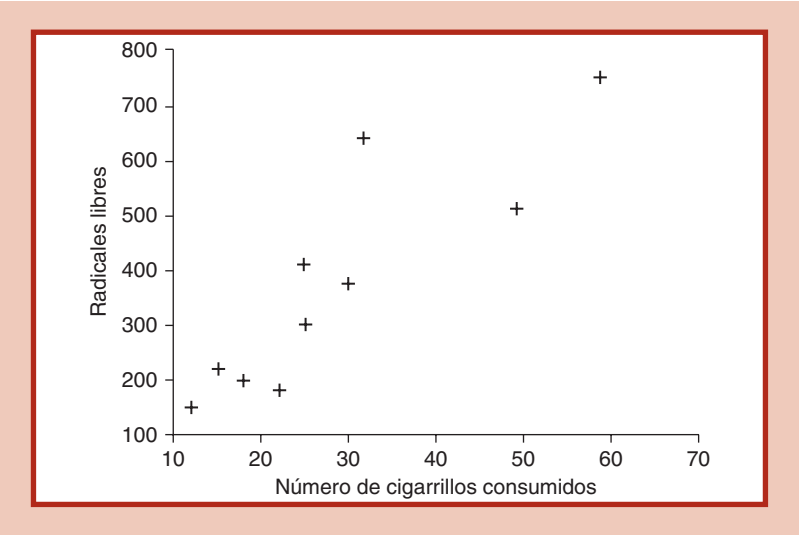


Figura 3.9 Consumo de cigarrillos frente a número de radicales libres.

Tabla 3.4 Pulsaciones por minuto y años de escolarización completados

	Persona									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Años de escolarización	12	16	13	18	19	12	18	19	12	14
Pulsaciones	73	67	74	63	73	84	60	62	76	71

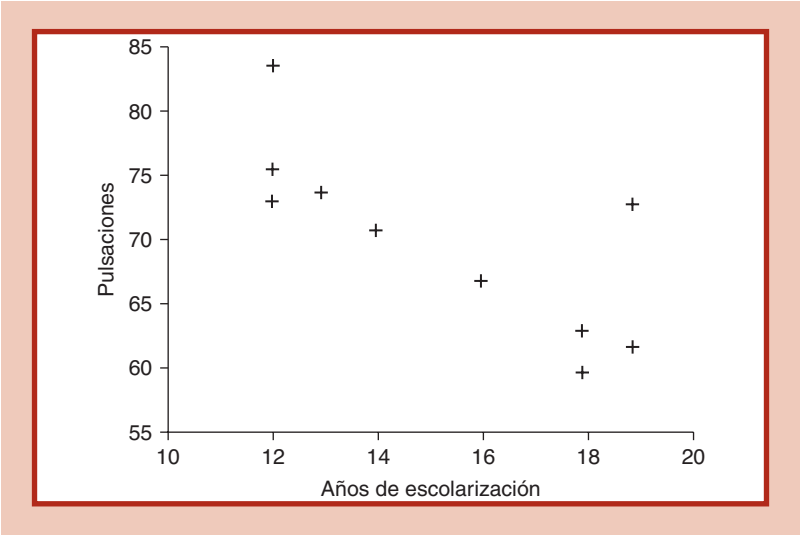


Figura 3.10 Diagrama de dispersión, de los años de escolarización y las pulsaciones por minutos.

Para obtener un estadístico que se pueda utilizar para medir la asociación entre los valores individuales de un conjunto de datos apareados, supongamos que los pares de valores del conjunto de datos son (x_i, y_i), $i = 1, \ldots, n$. Denotemos por $\bar{x}$ e $\bar{y}$ las medias muestrales de los valores x y de los valores y, respectivamente. Para cada par de valores i, consideremos $x_i - \bar{x}$, la desviación de su valor x de la media muestral, e $y_i - \bar{y}$, la desviación de su valor y de la media muestral. Ahora bien, si x_i es un valor x grande, será mayor que la media de todos los valores x y, por consiguiente, la desviación $x_i - \bar{x}$ será positiva. De igual forma, si x_i es un valor x pequeño, la desviación $x_i - \bar{x}$ será negativa. Puesto que lo mismo ocurre con las desviaciones y, se puede concluir lo siguiente:

Cuando los valores grandes de la variable x tienden a estar asociados con los valores grandes de la variable y, y si los valores pequeños de la variable x tienden a estar asociados con los valores pequeños de la variable y, los signos, positivos o negativos, de $x_i - \bar{x}$ e $y_i - \bar{y}$ de tienden a coincidir.

Ahora bien, si $x_i - \bar{x}$ e $y_i - \bar{y}$ tienen el mismo signo (positivo o negativo), su producto $(x_i - \bar{x})(y_i - \bar{y})$ será positivo. Por consiguiente, cuando los valores grandes de x tienden a estar asociados con los valores grandes de y, y si los valores pequeños de x tienden a estar asociados con valores pequeños de y, entonces $\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$ tenderá a tomar un valor positivo elevado.

La misma lógica implica que, cuando los valores grandes en una de las variables tienden a presentarse junto con los valores pequeños en la otra, los signos de $x_i - \bar{x}$ e $y_i - \bar{y}$ serán opuestos y, en consecuencia, $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ tomará un valor negativo elevado.

Para determinar qué significa que $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ tome un valor "elevado", se estandarizará esta suma y se dividirá por $n - 1$; después se dividirá entre el producto de las dos desviaciones típicas muestrales. El estadístico resultante se conoce con el nombre de *coeficiente de correlación muestral*.

Definición

Denotemos por s_x y s_y las desviaciones típicas muestrales de los valores x y de los valores y, respectivamente. El *coeficiente de correlación muestral*, representado por r, de los pares de datos (x_i, y_i), $i = 1, \ldots, n$, se define por

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Cuando $r > 0$, se dice que los pares de datos muestrales están *correlacionados positivamente*; y cuando $r < 0$, se dice que están *correlacionados negativamente*.

A continuación se listan algunas de las propiedades del coeficiente de correlación muestral.

1. El coeficiente de correlación muestral siempre está comprendido entre –1 y +1.
2. El coeficiente de correlación muestral r será igual a +1 si, para alguna constante a, se verifica que

   $$y_i = a + bx_i \quad \text{para } i = 1, \ldots, n$$

   donde b es una constante positiva.

3. El coeficiente de correlación muestral r será igual a –1 si, para alguna constante a, se verifica que

   $$y_i = a + bx_i \quad \text{para } i = 1, \ldots, n$$

   donde b es una constante negativa.

4. Si r es el coeficiente de correlación muestral para los datos x_i, y_i, $i = 1, \ldots, n$, para cualquiera de las constantes a, b, c, d, el coeficiente de correlación para los datos

$$a + bx_i, \; c + dy_i \quad i = 1, \ldots, n$$

coincide con r, en el caso de que b y d tengan el mismo signo (es decir, si $bd \geq 0$).

La propiedad 1 indica que el coeficiente de correlación muestral r siempre está entre -1 y $+1$. La propiedad 2 refleja que r será igual a $+1$ si los datos apareados están alineados (es decir, si existe una relación *lineal* entre ellos), de forma que los valores grandes de y se corresponden con valores grandes de x. La propiedad 3 indica que r es igual a -1 cuando existe una relación lineal entre los pares de valores, para la que los valores grandes de y están unidos a los valores pequeños de x. La propiedad 4 establece que el valor de r se mantiene invariable cuando se añade una constante a cada valor de la variable x (o a cada valor de la variable y) o cuando cada valor de la variable x (o a cada valor de la variable y) se multiplica por una constante positiva. Esta propiedad implica que r no depende de las unidades en que se miden los datos. Por ejemplo, el coeficiente de correlación muestral para los pesos y las alturas de cierto número de personas no depende de si las alturas se miden en pies o en pulgadas o de si los pesos se miden en libras o kilogramos. De igual forma, si uno de los valores de cada par es la temperatura, el coeficiente de correlación muestral es idéntico tanto si la temperatura se mide en grados Fahrenheit como si se mide en grados Celsius.

Desde un punto de vista de eficiencia computacional, la siguiente fórmula del coeficiente de correlación resulta ser apropiada.

Fórmula computacional para r

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}}$$

Ejemplo 3.22 La siguiente tabla muestra los consumos per cápita en Estados Unidos de leche entera y de leche desnatada durante tres años distintos.

	Consumo per cápita (en galones)		
	1980	**1984**	**1987**
Leche entera (x)	17,1	14,7	12,8
Leche desnatada (y)	10,6	11,5	13,2

Fuente: Departamento de Agricultura de Estados Unidos, *Consumo de alimentos, precios y gastos.*

Encuentre el coeficiente de correlación muestral para los datos dados.

Solución Para hacer que los cálculos sean más sencillos, empecemos restando 12,8 de cada valor x y restando 10,6 de cada valor y. Esto conduce al nuevo conjunto de datos:

	i		
	1	**2**	**3**
x_i	4,3	1,9	0
y_i	0	0,9	2,6

Ahora bien,

$$\bar{x} = \frac{4{,}3 + 1{,}9 + 0}{3} = 2{,}0667$$

$$\bar{y} = \frac{0 + 0{,}9 + 2{,}6}{3} = 1{,}1667$$

$$\sum_{i=1}^{3} x_i y_i = (1{,}9)(0{,}9) = 1{,}71$$

$$\sum_{i=1}^{3} x_i^2 = (4{,}3)^2 + (1{,}9)^2 = 22{,}10$$

$$\sum_{i=1}^{3} y_i^2 = (0{,}9)^2 + (2{,}6)^2 = 7{,}57$$

De donde,

$$r = \frac{1{,}71 - 3(2{,}0667)(1{,}1667)}{\sqrt{[22{,}10 - 3(2{,}0667)^2][7{,}57 - 3(1{,}1667)^2]}} = -0{,}97$$

Así pues, nuestros tres pares de datos muestran que existe una correlación negativa muy fuerte entre los consumos de leche entera y los de leche desnatada.

Para conjuntos de datos pequeños, tales como el del ejemplo 3.22, el coeficiente de correlación muestral puede obtenerse fácilmente a mano. Sin embargo, para conjuntos de datos grandes, su cálculo resulta tedioso y es conveniente usar una calculadora o un software estadístico. ■

Ejemplo 3.23 Calcule el coeficiente de correlación muestral para los datos de la tabla 3.3, en la que se relacionan los consumos de cigarrillos con el número de radicales libres en el interior de los pulmones de varios fumadores.

Solución El número de pares de datos es 10, cuyos valores son los siguientes:

18, 202
32, 644
25, 411
60, 755
12, 144
25, 302
50, 512

15, 223
22, 183
30, 375

Tras los cálculos pertinentes se llega a que el coeficiente de correlación muestral es 0,8759639. ■

El alto valor de este coeficiente de correlación muestral indica que existe una fuerte correlación positiva entre el consumo de cigarrillos de una persona y el número de radicales libres en el interior de sus pulmones.

Ejemplo 3.24 Calcule el coeficiente de correlación muestral para los datos de la tabla 3.4, donde se relacionan el número de pulsaciones por minuto de una persona con el número de años de escolarización que ha completado.

Solución Los pares de valores son los siguientes:

12, 73
16, 67
13, 74
18, 63
19, 73
12, 84
18, 60
19, 62
12, 76
14, 71

El coeficiente de correlación muestral es $-0{,}763803$.

El alto valor negativo de este coeficiente de correlación muestral indica que, para los datos en cuestión, un alto número de pulsaciones tiende a estar asociado a un bajo número de años de escolarización, y que un valor reducido en el número de pulsaciones tiende a corresponderse con un elevado número de años de escolarización. ■

El valor absoluto del coeficiente de correlación muestral r (esto es, $|r|$, su valor sin considerar el signo) es una medida de la fuerza de la relación lineal entre los valores x e y de cada par. Un valor de $|r|$ igual a 1 indica que existe una relación lineal perfecta; esto es, existe una recta que pasa por todos los puntos (x_i, y_i), $i = 1, \ldots, n$. Un valor de $|r|$ próximo a 0,8 indica que la relación lineal es relativamente fuerte; aunque no existe ninguna recta que pase a través de todos los puntos observados, existe una recta que pasa "cerca" de todos ellos. Un valor de $|r|$ próximo a 0,3 significa que la relación lineal es relativamente débil. El signo de r proporciona el sentido de la relación. Es positivo cuando la relación lineal es tal que los valores pequeños de y tienden a estar asociados con los valores pequeños de x, y cuando los valores grandes de y tienden a estar asociados con los valores igualmente grandes de x (por consiguiente, la relación lineal apunta hacia arriba); y es negativo cuando los valores grandes de y tienden a aparecer junto con los valores pequeños de x, y los valores pequeños de y tienden a aparecer junto con los valores grandes de x (por tanto, en este caso, la relación lineal apunta hacia abajo). En la figura 3.11 se reflejan los diagramas de dispersión de varios conjuntos de datos con distintos valores de r.

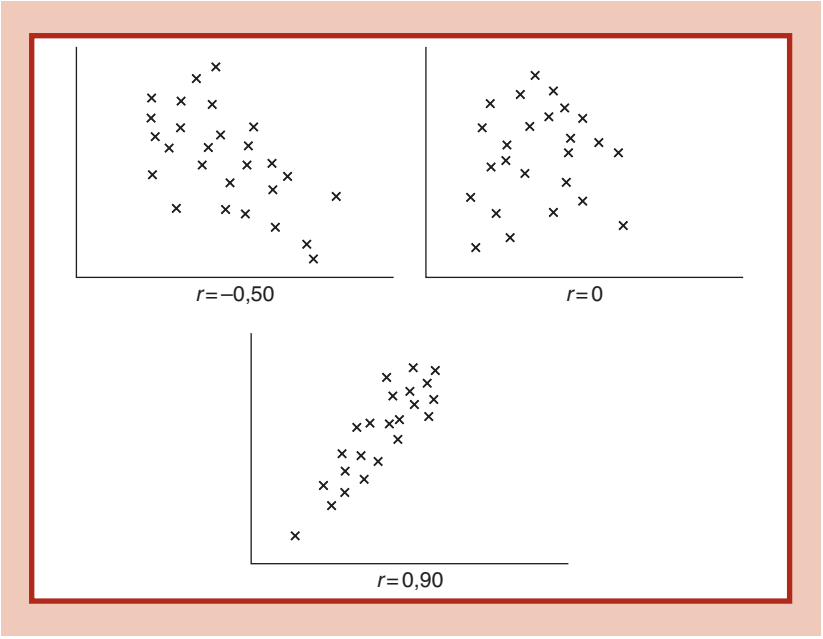


Figura 3.11 Coeficientes de correlación muestral.

(Bettmann)

Francis Galton

Perspectiva histórica

El desarrollo del coeficiente de correlación muestral y de su utilidad necesitó los esfuerzos de cuatro grandes estadísticos. La idea original fue de Francis Galton, quien estaba intentando estudiar las leyes de la herencia desde un punto de vista cuantitativo. Por este motivo, él quería ser capaz de cuantificar el grado en el que las características de un descendiente se relacionaban con las de sus padres. Ello le condujo a definir una forma de coeficiente de correlación muestral que difiere en cierta manera de la que se utiliza actualmente. Aunque originariamente pretendió utilizarlo para evaluar la influencia de la herencia de un padre sobre su descendencia, más tarde Galton se dio cuenta de que en realidad el coeficiente de correlación muestral era un método para evaluar la interrelación existente entre cualquier par de variables.

Aunque Francis Galton es considerado como el fundador de la Biometría –el análisis cuantitativo de la Biología–, Karl Pearson fue la figura más reconocida dentro de este ámbito, al menos con posterioridad a 1900. En ese año, la Real Sociedad de Londres aprobó una resolución en la que se indicaba que no se aceptarían más artículos que aplicaran las matemáticas a los estudios de Biología, y Pearson, con la ayuda financiera de Galton, fundó la revista estadística *Biometrika*, que todavía se edita hoy en día. La forma en que actualmente se utiliza el coeficiente de correlación muestral (que se ha presentado en este capítulo) se debe a Karl Pearson, por ello originalmente se conoció como *coeficiente de correlación del momento producto de Pearson*.

Las probabilidades asociadas a los posibles valores del coeficiente de correlación muestral r cuando los pares de datos provienen de poblaciones normales se deben a William Gosset. Sin embargo, en sus cálculos hubo ciertos errores técnicos que, posteriormente, fueron corregidos en un artículo de Ronald Fisher.

Problemas

1. Explique por qué el coeficiente de correlación muestral de los pares de datos

$$(121, 360), (242, 362), (363, 364)$$

es el mismo que el de los pares

$$(1, 0), (2, 2), (3, 4)$$

el cual, a su vez, coincide con el de los pares

$$(1, 0), (2, 1), (3, 2)$$

2. Calcule el coeficiente de correlación muestral para los pares de datos del problema 1.

Estadísticas en perspectiva

La correlación mide la asociación, no la causalidad

Los resultados del ejemplo 3.24 indican una fuerte correlación negativa entre los años de escolarización de los individuos y su número de pulsaciones cuando estaban en situación de descanso. Sin embargo, ello no implica que si aumenta el número de años de escolarización se reduzca directamente el número de pulsaciones por minuto. Es decir, el que los valores altos en el número de años de escolarización tiendan a estar asociados con los valores bajos en el número de pulsaciones no significa que los primeros sean la causa *directa* de los segundos. A menudo, la explicación de tal asociación se basa en un factor que no se ha tenido en cuenta, el que está relacionado con las dos variables que se consideren. En este ejemplo, podría ocurrir que una persona que hubiera estado escolarizada un alto número de años fuera más sensible a todo lo relacionado con el área de la salud y, en consecuencia, fuera más consciente de la importancia de hacer ejercicio y de tener buenos hábitos de alimentación; o quizá puede que no sea el conocimiento lo que establece la diferencia sino que, por el contrario, la gente con mayor educación acaba teniendo unos empleos que les permiten un mayor tiempo de ejercicio y mejores hábitos de nutrición. Probablemente, la fuerte correlación negativa encontrada entre los años de escolarización y el número de pulsaciones se deba a una combinación de estos y otros muchos factores subyacentes.

3. Los datos siguientes representan las puntuaciones obtenidas en un test de inteligencia (IQ) por 10 madres y por sus respectivas hijas mayores.

Puntuaciones de las madres	Puntuaciones de las hijas
135	121
127	131
124	112
120	115
115	99
112	118
104	106
96	89
94	92
85	90

(a) Dibuje un diagrama de dispersión.

(b) Haga una conjetura sobre el valor del coeficiente de correlación muestral r.

(c) Calcule el valor de r.

(d) ¿Qué conclusiones se pueden extraer acerca de la relación entre las puntuaciones de las madres y las de las hijas?

4. Los datos siguientes provienen de una muestra de 10 presos recientemente liberados que habían sido encarcelados por primera vez. Los datos incluyen el crimen cometido, su sentencia, y el tiempo real pasado en prisión.

Número	Crimen	Sentencia (en meses)	Tiempo en prisión (en meses)
1	Abuso de drogas	44	24
2	Falsificación	30	12
3	Abuso de drogas	52	26
4	Secuestro	240	96
5	Fraude de impuestos	18	12
6	Abuso de drogas	60	28
7	Robo	120	52
8	Desfalco	24	14
9	Robo	60	35
10	Robo	96	49

Dibuje un diagrama de dispersión de los tiempos de sentencia y de los tiempos reales en prisión. Calcule el coeficiente de correlación muestral. ¿Qué indica sobre la relación existente entre los tiempos sentenciados y los que se han cumplido realmente?

5. Con los datos del problema 4, determine el coeficiente de correlación muestral entre los tiempos de sentencia y las proporciones de estos tiempos que realmente se cumplieron.

¿Qué indica sobre la relación existente entre los tiempos sentenciados y dichas proporciones?

6. Los datos siguientes se refieren al número de adultos que están en prisión y de los que están en libertad condicional en los 12 Estados del occidente medio de Estados Unidos. Los datos están en miles de adultos.

Estado	En prisión	En libertad condicional
Illinois	18,63	11,42
Indiana	9,90	2,80
Iowa	2,83	1,97
Kansas	4,73	2,28
Michigan	17,80	6,64
Minnesota	2,34	1,36
Missouri	9,92	4,53
Nebraska	1,81	0,36
North Dakota	0,42	0,17
Ohio	20,86	6,51
South Dakota	1,05	0,42
Wisconsin	5,44	3,85

(a) Dibuje un diagrama de dispersión.

(b) Determine el coeficiente de correlación muestral entre el número de adultos en prisión y el número de adultos en libertad condicional.

(c) Rellene la palabra que falta. Los Estados que tienen un alto número de adultos en prisión tienden a tener un _____ número de adultos en libertad condicional.

7. Los siguientes datos relacionan el número de juicios criminales en varias ciudades de Estados Unidos y el porcentaje de sentencias de culpabilidad resultantes de ellos.

Ciudad	Porcentaje de juicios con sentencias de culpabilidad	Número de juicios criminales
San Diego, CA	73	11 534
Dallas, TX	72	14 784
Portland, OR	62	3 892
Chicago, IL	41	35 528
Denver, CO	68	3 772
Philadelphia, PA	26	13 796
Lansing, MI	68	1 358
St. Louis, MO	63	3 649
Davenport, IA	60	1 312
Tallahassee, FL	50	2 879
Salt Lake City, UT	61	2 745

Determine el coeficiente de correlación muestral entre el número de juicios criminales y el porcentaje de sentencias de culpabilidad. ¿Qué se puede decir sobre el grado de asociación entre estas dos variables que se han considerado?

8. Los siguientes datos relacionan los consumos per cápita de leche entera y de leche desnatada en Estados Unidos durante los años comprendidos entre 1980 y 1987, con la exclusión de 1981. (Algunos de estos datos fueron utilizados en el ejemplo 3.22.)

	Consumos (en galones)						
	1980	**1982**	**1983**	**1984**	**1985**	**1986**	**1987**
Leche entera	17,1	15,6	15,2	14,7	14,3	13,4	12,8
Leche desnatada	10,6	10,8	11,1	11,1	12,1	12,8	13,2

Fuente: Consumo de alimentos, precios y gastos.

Calcule el coeficiente de correlación muestral para los consumos de leche entera y de leche desnatada en los años citados.

9. Los siguientes datos muestran las rentas monetarias per cápita, en dólares, para 12 ciudades de Estados Unidos en los años 1979 y 1985.

Ciudad	**Renta en 1979**	**Renta en 1985**
New York	7 271	11 188
Baltimore	5 877	8 647
Denver	8 553	12 490
Austin	7 368	11 633
Cincinnati	6 874	10 247
Omaha	7 714	12 886
Detroit	6 215	8 852
Memphis	6 466	9 362
Milwaukee	7 029	9 765
St. Louis	5 877	8 799
Charlotte	7 952	12 259
Buffalo	5 929	8 840

Calcule el coeficiente de correlación muestral para las rentas per cápita de estas ciudades en 1979 y en 1985.

10. Los siguientes datos muestran el número de médicos y dentistas, por 100 000 habitantes, en Estados Unidos durante seis años diferentes

	1980	**1981**	**1982**	**1983**	**1985**	**1986**	**2001**
Médicos	211	217	222	228	237	246	253
Dentistas	54	54	55	56	57	57	59

Fuente: *Estadísticas de recursos sanitarios, anuario.*

(a) Compruebe si el número de médicos y el número de dentistas en los años citados están correlacionados positivamente.

(b) ¿Se puede pensar que un valor elevado en una de las dos variables causa por sí mismo un elevado valor en la otra? Si la respuesta es negativa, ¿cómo se podría explicar la correlación positiva existente?

En la tabla siguiente se incluyen las tasas de mortalidad, por una serie de causas seleccionadas, en diferentes países. Esta tabla será utilizada en los problemas del 11 al 13.

Tasas de mortalidad por 100 000 habitantes para las causas y los países seleccionados

				Neoplasia maligna de				
País	**Año**	**Enfermedad isquémica de corazón**	**Enfermedad cerebro-vascular**	**Pulmón, tráquea, bronquios**	**Estómago**	**Pecho (mujeres)**	**Bronquitis, enfisema, asma**	**Enfermedades de hígado y cirrosis crónicas**
Estados Unidos	1984	218,1	60,1	52,7	6,0	31,9	8,3	12,9
Alemania Occidental	1986	159,5	100,4	34,6	18,3	32,6	26,1	19,3
Australia	1985	230,9	95,6	41,0	10,1	30,0	16,9	8,7
Austria	1986	155,1	133,2	34,3	20,7	31,6	22,3	26,6
Bélgica	1984	120,6	95,0	55,9	14,7	36,8	22,6	12,4
Bulgaria	1985	245,9	254,5	30,6	24,2	21,5	28,6	16,2
Canadá	1985	200,6	57,5	50,6	9,0	34,5	9,7	10,1
Checoslovaquia	1985	289,4	194,3	51,3	22,4	27,3	33,8	19,6
Dinamarca	1985	243,8	73,4	52,2	10,9	39,7	37,1	12,2
España	1981	79,0	133,9	26,0	19,7	19,0	19,1	23,3
Finlanda	1986	259,8	105,0	36,4	17,3	23,9	19,8	8,8
Francia	1985	76,0	79,7	32,2	10,8	27,1	11,7	22,9
Holanda	1985	164,6	71,1	56,3	15,6	38,2	17,8	5,5
Hungría	1986	240,1	186,5	55,0	25,9	31,2	43,8	42,1
Italia	1983	128,9	121,9	42,1	23,9	28,9	30,9	31,5
Japón	1986	41,9	112,8	24,9	40,7	8,1	12,2	14,4
Noruega	1985	208,5	88,6	26,3	14,4	25,9	18,2	6,9
Nueva Zelanda	1985	250,5	98,4	42,0	11,2	37,7	25,8	4,8
Polonia	1986	109,4	75,3	47,2	24,2	21,1	33,4	12,0
Portugal	1986	76,6	216,4	18,7	26,5	22,6	17,8	30,0
Reino Unido:								
Escocia	1986	288,0	128,4	68,7	14,9	41,2	14,8	7,3
Inglaterra y Gales	1985	247,6	104,5	57,2	15,2	41,9	24,2	4,8
Suecia	1985	244,7	73,0	23,2	12,5	26,0	14,3	6,4
Suiza	1986	112,0	65,6	36,6	12,0	36,6	17,5	10,4

Fuente: Organización Mundial de la Salud, *Estadísticas de salud mundial*.

Si está ejecutando el programa 3-2 o se está usando algún paquete estadístico para resolver los problemas del 11 al 13, utilice todos los datos. Si está trabajando con una calculadora de mano, use sólo los datos referidos a los siete primeros países.

11. Encuentre el coeficiente de correlación muestral entre las tasas de mortalidad por enfermedad isquémica de corazón y por enfermedad de hígado crónica.

12. Encuentre el coeficiente de correlación muestral entre las tasas de mortalidad por cáncer de estómago y por cáncer de pecho en mujeres.

13. Encuentre el coeficiente de correlación muestral entre las tasas de mortalidad por cáncer de pulmón y por bronquitis, enfisema y asma.

14. En un famoso experimento, un investigador de la Universidad de Pittsburg solicitó la cooperación de los maestros de las escuelas públicas de Boston para conseguir un diente de leche de cada alumno. Después, serró todos los dientes que se habían recogido y determinó sus contenidos de plomo. Finalmente, hizo una representación gráfica de los contenidos de plomo frente a las puntuaciones de cada alumno en un test de inteligencia (IQ). Encontró una fuerte correlación negativa entre los contenidos de plomo y las puntuaciones citadas. Los periódicos resaltaron este hecho como una "prueba" de que las ingestiones de plomo producían un descenso en los niveles de inteligencia.

    (a) ¿Esta conclusión es necesariamente cierta?

    (b) Indique otras explicaciones posibles.

15. En un estudio reciente se encontró una fuerte correlación positiva entre los niveles de colesterol en adultos jóvenes y los tiempos que empleaban viendo la televisión.

    (a) ¿Era esperable tal resultado? ¿Por qué?

    (b) ¿Se puede pensar que ver televisión sea la causa de padecer mayores niveles de colesterol?

    (c) ¿Se puede pensar que tener niveles altos de colesterol hace que un joven adulto vea más televisión?

    (d) ¿Cómo se podría explicar el resultado del estudio?

16. Un análisis de los puntos conseguidos y de las faltas cometidas por los jugadores de baloncesto en la Conferencia del Pacífico estableció que existía una fuerte correlación positiva entre ambas variables. El analista difundió que este hecho prueba que los jugadores de baloncesto claramente ofensivos tienden a ser muy agresivos y que, en consecuencia, tienden a cometer un gran número de faltas. ¿Puede haber una explicación más simple para la correlación positiva encontrada? (*Sugerencia:* Piense en el número medio de minutos por juego que cada jugador está en pista.)

17. Un estudio publicado en octubre de 1993 en la revista *New England Journal of Medicine* encontró que la gente que tenía armas de protección en casa tenía tres veces más posibilidades de ser asesinados que aquellos que no tenían armas. ¿Prueba esto que las posibilidades de que un individuo sea asesinado se incrementan cuando decide comprar un arma para tenerla en casa? Explique su respuesta.

Términos clave

Estadístico: Magnitud numérica cuyo valor se puede determinar a partir de los datos.

Media muestral: Media aritmética de los valores de un conjunto de datos.

Desviación: Diferencia entre un valor de dato y la media muestral. Si x_i es el i-ésimo valor de dato y $\bar{x}$ es la media muestral, la diferencia $x_i - \bar{x}$ se denomina *desviación i-ésima.*

Mediana muestral: Valor central de un conjunto de datos ordenado. Para un conjunto de datos con n valores, la mediana muestral es el $(n + 1)/2$ valor menor, cuando n es impar; y es la media entre el $n/2$ y el $n/2 + 1$ menores valores, si n es par.

Percentil muestral de orden 100*p* por ciento: Valor de dato que cumple que al menos un $100p$ por ciento de los datos son menores o iguales que él y al menos un $100(1 - p)$ por ciento de los valores son mayores o iguales que él. Si existen dos valores de datos que cumplen estas condiciones, el percentil citado es igual a la media de ambos.

Primer cuartil: Percentil muestral de orden 25%.

Segundo cuartil: Percentil muestral de orden 50%, que también coincide con la mediana muestral.

Tercer cuartil: Percentil muestral de orden 75%.

Moda muestral: Valor de dato que ocurre con mayor frecuencia en un conjunto de datos.

Varianza muestral: Estadístico s^2, definido por

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}$$

Mide la media de las desviaciones al cuadrado.

Desviación típica muestral: Raíz cuadrada positiva de la varianza muestral.

Rango: Diferencia entre el mayor y el menor valor de dato.

Rango intercuartílico: Diferencia entre el tercer y el primer cuartil.

Conjunto de datos normal: Aquél cuyo histograma es simétrico con respecto a su intervalo central y que decrece a ambos lados de este intervalo siguiendo una forma acampanada.

Conjunto de datos asimétrico: Aquél cuyo histograma no es simétrico con respecto al intervalo de clase central. Se dice que es asimétrico por la derecha si su histograma presenta una cola alargada hacia la derecha, y se dice que es asimétrico por la izquierda si la cola alargada se sitúa hacia la izquierda.

Conjunto de datos bimodal: Aquél cuyo histograma presenta dos picos o chepas.

Coeficiente de correlación muestral: Para el conjunto de valores apareados x_i, y_i, $i = 1, \ldots, n$, se define por

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

donde $\bar{x}$ y s_x son, respectivamente, la media muestral y la desviación típica muestral de los valores x, y, de forma similar, se definen $\bar{y}$ y s_y. Un valor de r próximo a $+1$ indica que valores grandes de x tienden a estar apareados con valores grandes de y, y que valores pequeños de x tienden a estar apareados con valores pequeños de y. Un valor próximo a -1 indica que valores grandes de x tienden a estar apareados con valores pequeños de y, y que valores pequeños de x tienden a estar apareados con valores grandes de y.

Resumen

Se han visto tres estadísticos diferentes que describen el centro de un conjunto de datos: la media muestral, la mediana muestral y la moda muestral.

La media muestral de los datos $x_1 \ldots, x_n$ se define por

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

y es una medida del centro de los datos.

Si los datos vienen especificados mediante una tabla de frecuencias

Valor	Frecuencia
x_1	f_1
x_2	f_2
.	.
.	.
.	.
x_k	f_k

la media muestral de los $n = \sum_{i=1}^{k} f_i$ valores de datos puede expresarse como

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{n}$$

Una identidad de utilidad es

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

La mediana muestral es el valor central cuando los datos se encuentran ordenados de menor a mayor. Si existe un número par de datos, coincide con la media de los dos valores centrales. Es, también, una medida del centro de un conjunto de datos.

La moda muestral es el valor del conjunto de datos que ocurre con mayor frecuencia.

Supongamos que un conjunto de datos de tamaño n se ha ordenado de menor a mayor. Si np no es un entero, el percentil muestral de orden $100p$ por ciento se define como aquel valor que ocupa la posición que coincide con el menor entero que supera a np. Si np es un entero, el percentil muestral de orden $100p$ por ciento es la media entre los valores que ocupan las posiciones np y $np + 1$.

El percentil muestral de orden 25% es el *primer cuartil*. El percentil muestral de orden 50% (que coincide con la mediana muestral) se denomina *segundo cuartil*, y el percentil muestral de orden 75% se conoce como *tercer cuartil*.

La varianza muestral s^2 es una medida de la dispersión de los datos y se define por

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}$$

donde n es el tamaño del conjunto. Su raíz cuadrada positiva se denomina *desviación típica muestral*, y se mide en las mismas unidades que los datos.

La identidad

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}x_i^2 - n\bar{x}^2$$

resulta útil para calcular la varianza muestral con lápiz y papel o con una calculadora de mano.

El programa 3-1 permite computar la media muestral, la varianza muestral y la desviación típica muestral de cualquier conjunto de datos.

Otro estadístico que describe la dispersión de los datos es el *rango*, esto es, la diferencia entre el mayor y el menor valor de dato.

Los conjuntos de datos normales tienen su media muestral y su mediana muestral aproximadamente iguales. Sus histogramas son simétricos con respecto al intervalo central y tienen una forma acampanada.

El coeficiente de correlación muestral r mide el grado de asociación entre dos variables. Su valor está entre -1 y $+1$. Un valor de r próximo a $+1$ indica que cuando una de las variables es grande, la otra tiende a ser también grande, y cuando una de las variables es pequeña, la otra tiende igualmente a ser pequeña. Un valor de r cercano a -1 indica que cuando una de las variables es grande, la otra tiende a ser pequeña.

Un valor de $|r|$ grande indica la existencia de una fuerte asociación entre las dos variables. Asociación, sin embargo, no implica causalidad.

Problemas de repaso

1. Construya un conjunto de datos que sea simétrico con respecto a 0 y que contenga:
   (a) Cuatro valores distintos.
   (b) Cinco valores distintos.
   (c) En ambos casos, calcule la media muestral y la mediana muestral.
2. El siguiente gráfico de tallos y hojas refleja las presiones sanguíneas diastólicas de una muestra de 30 varones.

9	3, 5, 8,
8	6, 7, 8, 9, 9, 9
7	0, 1, 2, 2, 4, 5, 5, 6, 7, 8
6	0, 1, 2, 2, 3, 4, 5, 5
5	4, 6, 8

   (a) Calcule la media muestral .
   (b) Calcule la mediana muestral.
   (c) Obtenga la moda muestral.
   (d) Calcule la desviación típica muestral s.
   (e) ¿Los datos parecen ser aproximadamente normales?
   (f) ¿Qué proporción de valores de datos están comprendidos entre $\bar{x} + 2s$ y $\bar{x} - 2s$?
   (g) Compare la respuesta al apartado (f) con la proporción de datos entre ambos límites que se deduce de la regla empírica.

3. Los datos siguientes representan las edades medias de los residentes en cada uno de los 50 Estados de Estados Unidos.

29,3	27,7	30,4	31,1	28,5
32,1	28,0	31,3	26,6	25,8
25,9	33,0	31,5	30,0	28,4
24,9	31,6	26,6	25,4	29,2
29,3	27,9	31,8	31,5	30,3
28,5	29,3	26,6	31,2	32,1
31,4	30,1	27,0	28,5	27,6
28,9	29,4	30,5	31,2	29,4
29,3	30,1	28,8	27,9	30,4
32,3	30,4	25,8	27,1	26.9

   (a) Encuentre la mediana de estas edades.

   (b) Necesariamente, ¿ésta debe coincidir con la edad mediana de todos los habitantes de Estados Unidos?

   (c) Encuentre los cuartiles.

   (d) Encuentre el percentil muestral de orden 90%.

4. Utilice la tabla 3.2 (mostrada anteriormente) para completar la parte que falta en las frases siguientes:

   (a) Para que uno tenga una puntuación que esté dentro del 10% más alto de todos los estudiantes de Ciencias Físicas, debe ser de al menos ___.

   (b) Para que uno tenga una puntuación que esté dentro del 25% más alto de todos los estudiantes de Ciencias Sociales, debe ser de al menos ___.

   (c) Para que uno tenga una puntuación que esté dentro del 50% más bajo de todos los estudiantes de Medicina, debe ser de al menos ___.

   (d) Para que uno tenga una puntuación que esté dentro del 50% central de todos los estudiantes de Derecho, debe ser de al menos ___.

5. El número de crímenes violentos por 100 000 habitantes se muestra a continuación para cada uno de los 50 Estados de Estados Unidos. ¿Este conjunto de datos es aproximadamente normal?

Crímenes violentos por 100 000 habitantes, 2002

Estado	**Tasa de criminalidad**	**Rango de orden**
Estados Unidos	495	(X)
Alabama	444	21
Alaska	563	12
Arizona	553	13
Arkansas	424	22
California	593	10

Crímenes violentos por 100 000 habitantes, 2002 (*Continuación*)

Estado	Tasa de criminalidad	Rango de orden
Colorado	352	27
Connecticut	311	33
Delaware	599	9
Florida	770	2
Georgia	459	20
Hawaii	262	41
Idaho	255	42
Illinois	621	8
Indiana	357	26
Iowa	286	36
Kansas	377	24
Kentucky	279	38
Louisiana	662	6
Maine	108	48
Maryland	770	2
Massachusetts	484	18
Michigan	540	14
Minnesota	268	40
Mississippi	343	31
Missouri	539	15
Montana	352	27
Nebraska	314	32
Nevada	638	7
New Hampshire	161	47
New Jersey	375	25
New Mexico	740	4
New York	496	17
North Carolina	470	19
North Dakota	78	50
Ohio	351	29
Oklahoma	503	16
Oregon	292	34
Pennsylvania	402	23
Rhode Island	285	37
South Carolina	822	1
South Dakota	177	46
Tennessee	717	5
Texas	579	11
Utah	237	43

(*Continúa*)

Crímenes violentos por 100 000 habitantes, 2002 (*Continuación*)

Estado	Tasa de criminalidad	Rango de orden
Vermont	107	49
Virginia	291	35
Washington	345	30
West Virginia	234	44
Wisconsin	225	45
Wyoming	274	39

Observación: Los crímenes violentos se refieren a aquellos que fueron conocidos por la policía, incluyen asesinatos, secuestros forzados, robos y asaltos violentos. Cuando varios Estados comparten el mismo rango de orden, los siguientes rangos se omiten. Debido al redondeo de los datos, varios Estados pueden tener valores idénticos aunque un rango distinto.

6. Los datos siguientes representan los pesos de los recién nacidos en un hospital de una gran ciudad del este de Estados Unidos.

2,4, 3,3, 4,1, 5,0, 5,1, 5,2, 5,6, 5,8, 5,9, 5,9, 6,0, 6,1, 6,2, 6,3,
6,3, 6,4, 6,4, 6,5, 6,7, 6,8, 7,2, 7,4, 7,5, 7,5, 7,6, 7,6, 7,7, 7,8,
7,8, 7,9, 7,9, 8,3, 8,5, 8,8, 9,2, 9,7, 9,8, 9,9, 10,0, 10,3, 10,5

(a) Represéntelos gráficamente mediante un diagrama de tallos y hojas.

(b) Encuentre la media muestral $\bar{x}$.

(c) Encuentre la mediana muestral.

(d) Calcule la desviación típica muestral s.

(e) ¿Qué proporción de valores de datos están comprendidos entre $\bar{x} \pm 2s$?

(f) ¿Los datos parecen ser aproximadamente normales?

(g) Si su respuesta a (f) es sí, ¿qué proporción se estimaría para (e), si nos basamos en las respuestas a (b) y (d)?

*7. Sean a y b constantes. Demuestre que si $y_i = a + bx_i$, $i = 1, \ldots, n$, el coeficiente de correlación muestral, r, de los pares de datos x_i, y_i, $i = 1, \ldots, n$, viene dado por

(a) $r = 1$, si $b > 0$

(b) $r = -1$, si $b < 0$

(*Sugerencia:* Utilice la definición de r, y no su fórmula computacional.)

8. Los datos siguientes se han obtenido del libro *Investigaciones sobre la probabilidad de veredictos criminales y civiles,* publicado en 1837 por el matemático y probabilista francés Simeon Poisson. El libro enfatizaba las aplicaciones legales de la Probabilidad. Los datos se refieren al número de personas acusadas y condenadas por crímenes en Francia entre 1825 y 1830.

Año	Nº de acusados	Nº de condenados
1825	6652	4037
1826	6988	4348
1827	6929	4236
1828	7396	4551
1829	7373	4475
1830	6962	4130

(a) Determine la media muestral y la mediana muestral de los números de acusados.

(b) Determine la media muestral y la mediana muestral de los números de condenados.

(c) Determine la desviación típica muestral de los números de acusados.

(d) Determine la desviación típica muestral de los números de condenados.

(e) ¿Qué signo, positivo o negativo, se puede esperar que tenga el coeficiente de correlación muestral de las cifras de acusados y condenados?

(f) Determine el coeficiente de correlación muestral de los números de acusados y condenados.

(g) Determine el coeficiente de correlación muestral entre los números de acusados y los porcentajes de éstos que son condenados.

(h) Dibuje un diagrama de dispersión para los apartados (f) y (g).

(i) Haga una conjetura acerca del coeficiente de correlación muestral entre los números de condenados y los porcentajes de condenados sobre los acusados.

(j) Dibuje un diagrama de dispersión para las variables de (i).

(k) Determine el coeficiente de correlación muestral para las variables de (i).

9. Estudios recientes no han sido concluyentes sobre la posible conexión entre el consumo de café y la enfermedad coronaria de corazón. Un estudio indicó que los consumidores de grandes cantidades de café tenían mayores posibilidades de sufrir ataques de corazón que los consumidores moderados o los no consumidores, ¿prueba esto que el excesivo consumo de café incrementa el riesgo de sufrir un ataque de corazón? ¿Qué otras explicaciones son posibles?

10. Estudios recientes han indicado que las tasas de mortalidad de las personas casadas de mediana edad parecen ser menores que las de las personas solteras de mediana edad. ¿Significa esto que el matrimonio tiende a incrementar las longitudes de vida? ¿Qué otras explicaciones son posibles?

11. Un artículo del periódico *New York Times*, del 9 de junio de 1994, resaltaba un estudio en el que se mostraba que los años con bajos índices de inflación tendían a ser años con altos incrementos en la productividad media. En el artículo se argumentaba que este hecho apoyaba la tesis mantenida por la Reserva Federal en el sentido de que un bajo índice de inflación tiende a ocasionar un incremento en la productividad. Realmente, ¿se puede creer que el estudio proporciona una clara evidencia a favor de la tesis mantenida por la Reserva Federal? Explique la respuesta.

4

Probabilidad

La Probabilidad es la verdadera guía de la vida.

Cicerón, *De Natura*

Al principio de este capítulo se consideran experimentos cuyos resultados no se pueden predecir con certeza. Se definirán los sucesos de dichos experimentos. Posteriormente, se introducirá el concepto de probabilidad de un suceso, que coincide con la probabilidad de que el suceso contenga el resultado del experimento. También se dará una interpretación de la probabilidad de un suceso como el límite de la frecuencia relativa. Se tratarán las propiedades de las probabilidades. Después, se introducirá la probabilidad de un suceso condicionada a la ocurrencia de un segundo suceso. Finalmente, se verá qué se entiende por sucesos independientes.

4.1 Introducción

Se ha extraído una muestra representativa de 100 votantes para obtener información sobre la intención de voto en las próximas elecciones gubernamentales. Si 62 elementos de la muestra votan a favor del candidato republicano, ¿se puede concluir que la mayoría de los votantes están a favor de ese candidato? O, por el contrario, ¿es posible que, por *casualidad*, la muestra contenga una proporción mayor de votantes a favor de dicho candidato que

la proporción de votantes que le apoyan sobre el total de la población, y que, por tanto, el candidato demócrata sea el preferido de los electores?

Para responder a estas cuestiones, se necesita saber cuál es la probabilidad de que 62 personas de una muestra de 100 voten a favor de un candidato cuando, de hecho, éste no cuente con la mayoría de votantes en la población al completo. En realidad, como regla general, para realizar inferencias válidas sobre una población a partir de una muestra se necesita conocer la probabilidad de que ocurran ciertos sucesos bajo distintas circunstancias. La determinación de la verosimilitud, o la posibilidad, de que ocurra un suceso es el objetivo de la *Probabilidad*.

4.2 Espacio muestral y sucesos de un experimento

El término *probabilidad* se utiliza habitualmente en relación con la posibilidad de que ocurra un determinado suceso cuando se lleva a cabo un *experimento*, concebido éste en un sentido muy amplio. De hecho, para nosotros un *experimento* será *cualquier* proceso del que se deduzca una observación, o un *resultado*.

Con frecuencia estaremos interesados en experimentos cuyos resultados no sean, de antemano, predecibles con certeza. Y aunque el resultado del experimento no se conozca por adelantado, se supondrá que el conjunto de sus posibles resultados sí que es conocido. Este conjunto de todos los resultados posibles de un experimento se denominará *espacio muestral* y se denotará por S.

Definición

Un *experimento* es cualquier proceso que produzca una observación o *resultado*. El conjunto de todos los posibles resultados de un experimento se denomina *espacio muestral*.

Ejemplo 4.1 A continuación se muestran distintos ejemplos de experimentos y de sus espacios muestrales.

(a) Si el resultado de un experimento es el sexo de un descendiente, el espacio muestral será

$$S = \{g, b\}$$

donde la salida g indica que se trata de una hembra y b indica que se trata de un varón.

(b) Si el experimento consiste en el lanzamiento de dos monedas cuyos resultados sean cara o cruz, el espacio muestral será

$$S = \{(\text{C}, \text{C}), (\text{C}, \text{Z}), (\text{Z}, \text{C}), (\text{Z}, \text{Z})\}$$

El resultado es (C, C) si se obtienen caras en los dos lanzamientos, (C, Z) si el primer lanzamiento es cara y el segundo cruz, (Z, C) si el primero es cruz y el segundo cara, y (Z, Z) si en ambos lanzamientos se obtienen cruces.

(c) Si la salida de un experimento es el orden de llegada a la meta en una carrera en la que participan 7 caballos identificados con los números 1, 2, 3, 4, 5, 6 y 7, el espacio muestral será

$$S = \{\text{todas las permutaciones de } 1, 2, 3, 4, 5, 6, 7\}$$

La salida (4, 1, 6, 7, 5, 3, 2) significa que el caballo 4 llega en primer lugar, el caballo 1 llega en segundo lugar, y así sucesivamente.

(d) Considere un experimento que consiste en observar las caras resultantes en el lanzamiento de dos dados. Si el primer dado se identifica por 1 y el segundo por 2, el par de resultados de los dados 1 y 2 pueden representar los resultados de este experimento. Denotemos como (i, j) el que se obtenga i en el lanzamiento del dado 1 y j en el lanzamiento del dado 2. El espacio muestral será en ese caso

$$S = \{(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), (2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6), (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6), (4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6), (5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6), (6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)\}$$ ■

Cualquier conjunto de resultados del experimento se denomina *suceso*. Es decir, un suceso es un subconjunto del espacio muestral. Los sucesos se denotarán mediante las letras mayúsculas *A, B, C*, etc.

Ejemplo 4.2 En el ejemplo 4.1(a), si $A = \{g\}$, A será el suceso de que el descendiente sea una niña. De igual forma, si $B = \{b\}$, B será el suceso de que el descendiente sea un niño.

En el ejemplo 4.1(b), si $A = \{(C, C), (C, Z)\}$, A será el suceso de que resulte cara en el primero de los dos lanzamientos.

En el ejemplo 4.1(c), si

$$A = \{\text{todos los resultados de } S \text{ que comienzan con } 2\},$$

A será el suceso de que el caballo 2 gane la carrera.

En el ejemplo 4.1(d), si

$$A = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$$

A será el suceso de que la suma de los dos lanzamientos sea 7. ■

Definición

Cualquier conjunto de resultados de un experimento se denomina *suceso*. Los sucesos se denominarán *A, B, C,* etc. Se dice que un suceso A ocurre si el resultado está contenido en A.

Dados dos sucesos A y B, se define el nuevo suceso $A \cup B$, llamado *unión* de A y B, como aquel que incluye todos los resultados que están en A, en B, o en ambos.

En el ejemplo 4.1(a), si $A = \{g\}$ es el suceso de que el descendiente sea una niña y $B = \{b\}$ es el suceso de que el descendiente sea un niño, $A \cup B$. Es decir, $A \cup B$ coincide con el espacio muestral.

En el ejemplo 4.1(c), sea

$$A = \{\text{resultados de } S \text{ que comienzan con } 4\}$$

el suceso de que el caballo 4 gane la carrera, y sea

$$B = \{\text{resultados de } S \text{ cuyo segundo elemento sea } 2\}$$

el suceso de que el caballo 2 termine en segundo lugar. En este caso, $A \cup B$ es el suceso de que el caballo 4 gane, o de que el caballo 2 acabe en segundo lugar, o de que ambos ocurran simultáneamente.

Se puede llevar a cabo una útil representación gráfica de los sucesos mediante los *diagramas de Venn*. En éstos, el espacio muestral se identifica con todos los puntos de un rectángulo. Los sucesos de interés se indican sombreando distintas regiones del diagrama. La región coloreada de la figura 4.1 representa la unión de los sucesos A y B.

Dados dos sucesos cualesquiera A y B, la *intersección* de A y B consiste, por definición, en todos los resultados que están simultáneamente en A y en B. Esto es, ocurre la intersección si ocurren tanto A como B. $A \cap B$ denotará la intersección de A y B. La región coloreada de la figura 4.2 representa la intersección de los sucesos A y B.

En el ejemplo 4.1(b), si $A = \{(C, C), (C, Z)\}$ es el suceso consistente en obtener cara en el primero de dos lanzamientos sucesivos de una moneda y $B = \{(C, Z), (Z, Z)\}$ es el suceso de que el segundo lanzamiento sea cruz, se tiene que $A \cap B = \{(C, Z)\}$ es el suceso de obtener cara en el primer lanzamiento de la moneda y cruz en el segundo.

En el ejemplo 4.1(c), si A es el suceso de que el caballo 2 gane y B es el suceso de que gane el caballo 3, el suceso $A \cap B$ no contiene ningún resultado y, por tanto, no puede ocurrir. Un suceso que no contenga ningún resultado se denominará suceso *nulo*, y se designará

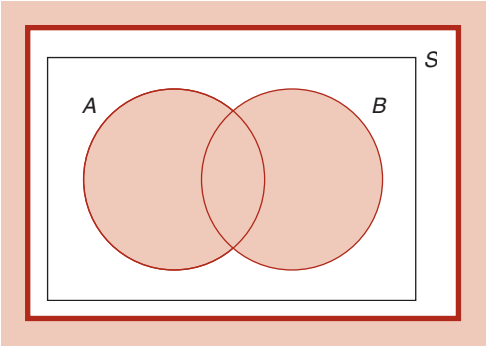


Figura 4.1 Diagrama de Venn: la región sombreada es $A \cup B$.

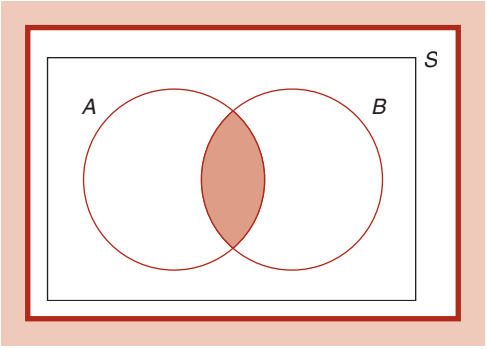


Figura 4.2 La región sombreada es $A \cap B$.

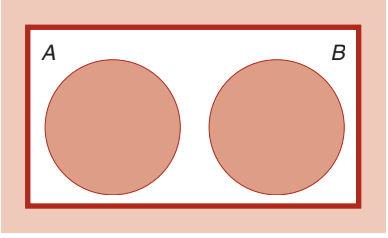


Figura 4.3 A y B son sucesos disjuntos.

por Ø. Si la intersección de A y B es el suceso nulo, se dirá que A y B son *disjuntos* o *mutuamente excluyentes*, puesto que ambos sucesos no pueden ocurrir simultáneamente. En el diagrama de Venn de la figura 4.3 están representados dos sucesos disjuntos.

Para cualquier suceso A, se define el suceso A^c, llamado *complementario de* A, como aquel que contiene todos los resultados del espacio muestral que no están en A. Es decir, A^c ocurrirá cuando no ocurra A, y viceversa. Si en el ejemplo 4.1(a), $A = \{g\}$ es el suceso de que el descendiente sea hembra, $A^c = \{b\}$ es el suceso de que el descendiente sea varón. Observe, además, que el complemento del espacio muestral es el conjunto nulo; es decir, $S^c = \text{Ø}$. La figura 4.4 muestra A^c, el complementario del suceso A.

Igualmente se pueden definir las uniones y las intersecciones de más de dos conjuntos. Por ejemplo, la unión de los sucesos A, B y C, que se escribirá $A \cup B \cup C$, contendrá todos los resultados que estén en A o en B o en C. Así pues, ocurrirá $A \cup B \cup C$ si ocurre al menos uno de estos tres sucesos. De igual forma, la intersección $A \cap B \cap C$ contendrá todos los resultados que estén en los tres sucesos A, B y C. Por consiguiente, la intersección ocurrirá sólo si ocurren todos los sucesos.

Se dice que los sucesos A, B y C son *disjuntos* si dos cualquiera de ellos no pueden ocurrir simultáneamente.

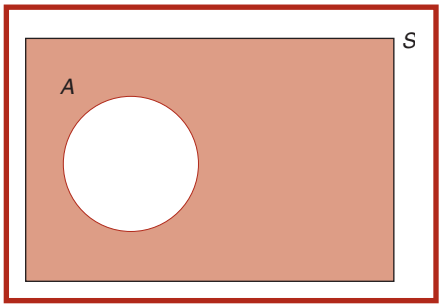


Figura 4.4 La región sombreada es A^c.

Problemas

1. Una caja contiene tres bolas: una roja, una azul y una amarilla. Considere el experimento de extraer una bola de la caja, devolverla a la caja y extraer una segunda bola.

   (a) ¿Cuál es el espacio muestral de este experimento?

   (b) ¿Cuál es el suceso de que la primera bola extraída sea amarilla?

   (c) ¿Cuál es el suceso de que la misma bola se extraiga dos veces?

2. Repita el problema 1 cuando la segunda bola se extrae sin haber reemplazado la primera bola.

3. Audrey y su amigo Charles deben elegir a qué universidad irán el próximo curso. Audrey ha sido admitida en las universidades de Michigan (MI), Reed (OR), San Jose State (CA), Yale (CT) y Oregon State (OR); entre paréntesis se indica el Estado de Estados Unidos en el que se encuentra cada una. Charles fue admitido en las universidades de Oregon State y San Jose State. Supongamos que el resultado del experimento sean las universidades a las que decidan ir Audrey y Charles.

   (a) Liste todos los resultados del espacio muestral S.

   (b) Liste los resultados del suceso de que Audrey y Charles elijan la misma universidad.

   (c) Liste los resultados del suceso de que Audrey y Charles elijan universidades distintas.

   (d) Liste los resultados del suceso de que Audrey y Charles elijan universidades del mismo Estado.

4. Un experimento consiste en el lanzamiento de una moneda tres veces, y en cada ocasión se anota si se obtiene cara o cruz.

   (a) ¿Cuál es el espacio muestral de este experimento?

   (b) ¿Cuál es el suceso consistente en que resulten más cruces que caras?

5. Los miembros de una familia han decidido pasar sus próximas vacaciones en Francia o en Canadá. Si van a Francia, pueden ir en avión o en barco. Si van a Canadá pueden ir en coche, en tren o en avión. Si el resultado del experimento consiste en el país y el tipo de desplazamiento elegidos, liste todos los puntos del espacio muestral. Liste, igualmente, todos los resultados del suceso A, consistente en que la familia vuele a su destino.

6. Los Yankees de Nueva York y los White Sox de Chicago van a jugar tres partidos este fin de semana. Suponiendo que se trata de tres partidos de eliminatoria (sin que se acepte el empate) y que se está interesado solamente en saber qué equipo gana cada partido, liste todos los resultados del espacio muestral S. Liste, también, todos los resultados de A, suceso consistente en que los Yankees ganen más partidos que los White Sox.

7. Sean $S = \{1, 2, 3, 4, 5, 6\}$, $A = \{1, 3, 5\}$, $B = \{4, 6\}$ y $C = \{1, 4\}$. Encuentre

   (a) $A \cap B$

   (b) $B \cup C$

   (c) $A \cup (B \cap C)$

   (d) $(A \cup B)^c$

   Observación: Las operaciones entre paréntesis se han de realizar antes. Por ejemplo, en (c) se debe determinar primero la intersección de B y C, y luego se ha de hacer la unión de A y ese conjunto.

8. Una cafetería ofrece un menú con tres platos. Se puede elegir un plato principal, un complemento y un postre. Las elecciones posibles de cada uno son las siguientes:

Menú	**Elección**
Plato principal	Pollo o filete
Complemento	Pasta, arroz o patata
Postre	Helado, gelatina o tarta de manzana

Cada individuo debe elegir un elemento de cada categoría.

   (a) Liste todos los resultados del espacio muestral S.

   (b) Si A es el suceso consistente en que se elija helado; liste los resultados de A.

   (c) Si B es el suceso consistente en que se elija pollo, liste los resultados de B.

   (d) Liste todos los resultados de $A \cap B$.

   (e) Si C es el suceso consistente en que se elija arroz, liste los resultados de C.

   (f) Liste todos los resultados del suceso $A \cap B \cap C$.

9. Un hospital clasifica a cada paciente según disponga o no de seguro médico y según su estado de salud, que puede ser catalogado como bueno, aceptable, serio o crítico. El administrador registra primero un 0 si el paciente no tiene seguro y un 1 si lo tiene, y después registra una de las letras *b, a, s* o *c*, según el estado en que se encuentra el

paciente. Por ejemplo, la codificación 1, *b* se refiere a un paciente con seguro y un estado de salud bueno. Considere que el experimento consiste en otorgar un código a un paciente nuevo.

(a) Liste el espacio muestral de este experimento.

(b) Indique cuál es el suceso de que el paciente esté en un estado serio o crítico y de que no tenga seguro médico.

(c) Indique cuál es el suceso de que el paciente esté en un estado bueno o aceptable.

(d) Especifique el suceso correspondiente a que el paciente disponga de seguro médico.

10. Los siguientes pares de sucesos E y F se refieren al mismo experimento. Diga en cada caso si E y F son sucesos disjuntos.

(a) Se lanza un dado. El suceso E es que salga un número par, y el suceso F es que salga un número impar.

(b) Se lanza un dado. El suceso E es que salga un 3, y el suceso F es que salga un número par.

(c) Se selecciona a una persona. El suceso E es que dicha persona haya nacido en Estados Unidos, mientras que el suceso F es que dicha persona sea un ciudadano de Estados Unidos.

(d) Se selecciona a un varón. El suceso E es que tenga una edad superior a 30 años, y el suceso F es que haya estado casado durante más de 30 años.

(e) Se selecciona a una mujer de una cola de personas que están esperando para matricular su coche. El suceso E es que el coche haya sido fabricado en Estados Unidos, mientras que el suceso F es que haya sido fabricado fuera de Estados Unidos.

11. Si A es el suceso de obtener un número par cuando se lanza un dado.

(a) Describa con palabras el suceso A^c.

(b) Describa con palabras el suceso $(A^c)^c$.

(c) En general, si A es un suceso cualquiera, ¿cuál es el suceso complementario de su complementario? Es decir, ¿cuál es el suceso $(A^c)^c$?

12. Se lanza un dado dos veces. Si los sucesos son: A, que la suma de los dos lanzamientos sea par; B, que resulte un *1* en el primer lanzamiento, y C que la suma de los dos lanzamientos sea 6. Describa los sucesos siguientes:

(a) $A \cap B$

(b) $A \cup B$

(c) $B \cap C$

(d) B^c

(e) $A^c \cap C$

(f) $A \cap B \cap C$

13. Sean A, B y C tres sucesos. Utilice los diagramas de Venn para representar que:

(a) Sólo ocurra A.

(b) Ocurran A y B, pero no C.

(c) Ocurra al menos uno de los tres sucesos.

(d) Ocurran al menos dos de los citados sucesos.

(e) Ocurran los tres sucesos simultáneamente.

4.3 Propiedades de la Probabilidad

Es un hecho empíricamente comprobado que, si se repite un experimento sucesivamente bajo las mismas condiciones, se verifica que, para cualquier suceso A, la proporción de resultados contenidos en A se aproxima a cierto valor a medida que el número de repeticiones aumenta. Por ejemplo, si se lanza una moneda sucesivamente, la proporción de lanzamientos en los que se obtiene cruz se aproxima a un valor a medida que el número de lanzamientos crece. Esta proporción, o frecuencia relativa, a largo plazo es lo que uno tiene en mente cuando se habla de la probabilidad de un suceso.

Consideremos un experimento cuyo espacio muestral sea S. Supongamos que para cada suceso A existe un número, denotado por $P(A)$ y llamado *probabilidad* del suceso A, que verifica las tres propiedades siguientes:

PROPIEDAD 1: Para cualquier suceso A, la probabilidad de A es un número comprendido entre 0 y 1. Esto es,

$$0 \leq P(A) \leq 1$$

PROPIEDAD 2: La probabilidad del espacio muestral S es 1. Es decir,

$$P(S) = 1$$

PROPIEDAD 3: La probabilidad de una unión de sucesos disjuntos es igual a la suma de las probabilidades de dichos sucesos. Por ejemplo, si A y B son disjuntos:

$$P(A \cup B) = P(A) + P(B)$$

El valor $P(A)$ representa la probabilidad de que el resultado del experimento esté contenido en el suceso A. La propiedad 1 establece que la probabilidad de que el resultado del experimento caiga dentro de A está comprendida entre 0 y 1. La propiedad 2 establece que, con probabilidad 1, el resultado del experimento es un elemento del espacio muestral S. La propiedad 3 establece que, si los sucesos A y B no pueden ocurrir simultáneamente, la probabilidad de que el resultado del experimento esté contenido bien en A o

bien en B es igual a la suma de la probabilidad de que esté en A y de la probabilidad de que esté en B.

Si se interpreta $P(A)$ como el límite de la frecuencia relativa de un suceso A, las condiciones establecidas se cumplen. La proporción de experimentos en los que A contenga el resultado será con seguridad un número comprendido entre 0 y 1. La proporción de experimentos en los que S contiene al resultado es 1, puesto que todos los resultados están contenidos en el espacio muestral S. Finalmente, si A y B no contienen resultados comunes, la proporción de experimentos cuyos resultados estén en A o en B es igual a la proporción de experimentos cuyos resultados estén en A más la proporción de experimentos cuyos resultados estén en B. Por ejemplo, si la proporción de lanzamientos de un par de dados cuyos resultados sumen 7 es 1/6 y la proporción de lanzamientos cuyos resultados sumen 11 es 1/18, la proporción de lanzamientos con una suma resultante igual a 7 o 11 es $1/6 + 1/18 = 2/9$.

Se pueden utilizar las propiedades 1, 2 y 3 para establecer algunos resultados generales relativos a las probabilidades. Por ejemplo, puesto que A y A^c son sucesos disjuntos cuya unión es el espacio muestral al completo, se puede escribir

$$S = A \cup A^c$$

A partir de las propiedades 2 y 3 se deduce lo siguiente:

$$\begin{aligned} 1 &= P(S) && \text{por la propiedad 2} \\ &= P(A \cup A^c) \\ &= P(A) + P(A^c) && \text{por la propiedad 3} \end{aligned}$$

Por consiguiente, se ve que

$$P(A^c) = 1 - P(A)$$

Así, la probabilidad de que el resultado de un experimento no esté contenido en A es 1 menos la probabilidad de que sí esté contenido en A. Por ejemplo, si la probabilidad de obtener cara en el lanzamiento de una moneda es de 0,4, la probabilidad de obtener cruz será de 0,6.

La fórmula siguiente relaciona la probabilidad de la unión de los sucesos A y B, no necesariamente disjuntos, con $P(A)$, $P(B)$ y la probabilidad de la intersección de A y B. Se la conoce habitualmente como *regla de adición de la probabilidad*.

Regla de adición

Para los sucesos A y B,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Para ver por qué la regla de adición es cierta, observe que $P(A \cup B)$ es la probabilidad de todos los resultados que se encuentran en A o en B. Por otro lado, $P(A) + P(B)$ es la probabilidad de todos los resultados que están en A más la probabilidad de todos los resultados

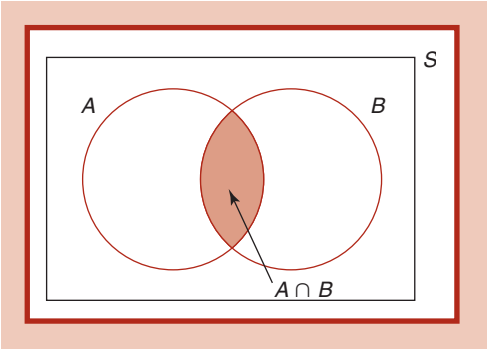


Figura 4.5 $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

que están en B. Puesto que todo suceso que esté tanto en A como en B se cuenta dos veces en $P(A) + P(B)$ y sólo una en $P(A \cup B)$ (véase la figura 4.5), se sigue que

$$P(A) + P(B) = P(A \cup B) + P(A \cap B)$$

Si se resta $P(A \cap B)$ en los dos términos de la ecuación anterior se obtiene la regla de adición.

El ejemplo 4.3 ilustra la utilidad de esta regla.

Ejemplo 4.3 Una tienda acepta cualquiera de las tarjetas de crédito American Express y VISA. Un 22% de sus clientes llevan la American Express, un 58% disponen de VISA y un 14% disponen de ambas. ¿Cuál es la probabilidad de que un cliente tenga al menos una de las dos tarjetas citadas?

Solución A denota el suceso de que el cliente disponga de la tarjeta American Express, y B denota el suceso de que disponga de la tarjeta VISA. Con la información citada se llega a

$$P(A) = 0{,}22 \quad P(B) = 0{,}58 \quad P(A \cap B) = 0{,}14$$

Por la regla de adición, la probabilidad pedida $P(A \cup B)$ es

$$P(A \cup B) = 0{,}22 + 0{,}58 - 0{,}14 = 0{,}66$$

Esto es, el 66% de los clientes del establecimiento poseen al menos una de las dos tarjetas aceptadas. ■

Para ilustrar la interpretación de la probabilidad como límite de la frecuencia relativa, se han simulado 10 000 lanzamientos de una moneda perfectamente simétrica. La tabla 4.1 muestra los números totales de caras obtenidas en los 10, 50, 100, 500, 2000, 6000, 8000 y 10 000 lanzamientos, junto con las correspondientes proporciones de caras. Observe que la proporción de caras se va aproximando cada vez más a 0,5, a medida que el número de lanzamientos crece.

Tabla 4.1 10 000 lanzamientos de una moneda simétrica.

n	Número de caras en los n primeros lanzamientos	Número de cruces en los n primeros lanzamientos	Proporción de caras en los n primeros lanzamientos
10	3	7	0,3
50	21	29	0,42
100	46	54	0,46
500	248	252	0,496
2 000	1 004	996	0,502
6 000	3 011	2 989	0,5018
8 000	3 974	4 026	0,4968
10 000	5 011	4 989	0,5011

Tabla 4.2 10 000 lanzamientos de un dado simétrico

	i					
	1	**2**	**3**	**4**	**5**	**6**
Frecuencia de ocurrencia	1724	1664	1628	1648	1672	1664
Frecuencia relativa	0,1724	0,1664	0,1628	0,1648	0,1672	0,1664

Observación: 1/6 = 0,166667.

La tabla 4.2 muestra los resultados de 10 000 lanzamientos simulados de un dado perfectamente simétrico.

Problemas

1. Suponga que el espacio muestral de un experimento es

$$S = \{1, 2, 3, 4, 5, 6\}$$

Si A_i es el suceso que consiste en obtener el resultado individual i, y si suponemos que

$$P(A_1) = 0{,}1 \qquad P(A_4) = 0{,}15$$
$$P(A_2) = 0{,}2 \qquad P(A_5) = 0{,}1$$
$$P(A_3) = 0{,}15 \qquad P(A_6) = 0{,}3$$

Es decir, la probabilidad de obtener el resultado 1 es 0,1; la de obtener el resultado 2 es 0,2; la de obtener el resultado 3 es 0,15, y así sucesivamente. Los sucesos E, F y G son los siguientes:

$$E = \{1, 3, 5\} \qquad F = \{2, 4, 6\} \qquad G = \{1, 4, 6\}$$

Perspectiva histórica

(Bettmann)

Pierre Fermat

(Bettmann)

Blaise Pascal

El tratamiento numérico de la probabilidad es relativamente reciente. De hecho, durante mucho tiempo a lo largo de la historia se creyó que todo lo que ocurría en la vida venía determinado por unas fuerzas superiores a la capacidad de entendimiento de las personas. Fue durante la primera mitad del siglo XVII, casi al final del Renacimiento, cuando la gente sintió curiosidad por el mundo y por las leyes que lo gobernaban. Entre esos curiosos se encontraban algunos amantes del juego. Un grupo de jugadores italianos, incapaces de dar solución a ciertas cuestiones relacionadas con lanzamientos de dados, contactaron con el famoso científico Galileo. Éste, aunque andaba ocupado en otros temas, se interesó por los problemas que le planteaban, y no sólo encontró soluciones a esos problemas sino que, además, escribió un pequeño tratado sobre los juegos de azar.

Unos años más tarde, sucedió una historia similar en Francia, donde residía un jugador conocido como Chevalier de Mere. Este apasionado jugador y matemático amateur conocía al brillante matemático Blaise Pascal, y le pidió que le ayudara a resolver algunos de sus problemas de juegos más complejos. Uno de ellos era el *problema de los puntos*, relativo a la división equitativa de la apuesta si dos jugadores decidían interrumpir el juego de azar antes de que hubiera finalizado. A Pascal este problema le pareció particularmente intrigante y, en 1654, se lo comentó por carta al matemático Pierre Fermat. La correspondencia entre ellos no sólo condujo a la solución de aquel problema en particular, sino que también sirvió de marco para resolver otros muchos problemas relacionados con los juegos de azar. Por ese motivo, algunos consideran la correspondencia entre ellos como el nacimiento de la Probabilidad, pues estimuló su interés entre algunos de los matemáticos más prominentes de la época. Por ejemplo, el joven genio holandés Lugwig Huyghens viajó a París para profundizar en esa nueva disciplina, y la actividad en esta área creció rápidamente.

Encuentre:

(a) $P(E)$, $P(F)$, $P(G)$ (b) $P(E \cup F)$

(c) $P(E \cup G)$ (d) $P(F \cup G)$

(e) $P(E \cup F \cup G)$ (f) $P(E \cap F)$

(g) $P(F \cap G)$ (h) $P(E \cap G)$

(i) $P(E \cap F \cap G)$

2. Si A y B son sucesos disjuntos para los que $P(A) = 0{,}2$ y $P(B) = 0{,}5$, encuentre:

(a) $P(A^c)$

(b) $P(A \cup B)$

(c) $P(A \cap B)$

(d) $P(A^c \cap B)$

3. La fenilcetonuria es una enfermedad genética que ocasiona un retraso mental. Aproximadamente, uno de cada 10 000 recién nacidos vivos la padecen. ¿Cuál es la probabilidad de que el próximo bebé que nazca en un hospital de Houston la padezca?

4. Una determinada persona encuentra tres semáforos en su trayecto al trabajo. Suponga que los valores siguientes reflejan las probabilidades del número total de semáforos que encuentra en rojo y en los que, por tanto, se debe parar:

$$P(0 \text{ semáforos en rojo}) = 0{,}14$$

$$P(1 \text{ semáforos en rojo}) = 0{,}36$$

$$P(2 \text{ semáforos en rojo}) = 0{,}34$$

$$P(3 \text{ semáforos en rojo}) = 0{,}16$$

   (a) ¿Cuál es la probabilidad de que, de camino al trabajo, se tenga que parar al menos una vez?

   (b) ¿Cuál es la probabilidad de que se tenga que parar en más de dos semáforos?

5. Si A y B son sucesos disjuntos, ¿es posible que ocurra lo siguiente?

$$P(A) + P(B) = 1{,}2$$

   ¿Y si A y B no son sucesos disjuntos?

6. Si la probabilidad de sacar un rey con una baraja de cartas de pinacle es 1/6 y la probabilidad de sacar un as es 1/6, ¿cuál es la probabilidad de sacar un rey o un as?

7. Suponga que la demanda de árboles navideños en un comercio es:

1100	con probabilidad 0,2
1400	con probabilidad 0,3
1600	con probabilidad 0,4
2000	con probabilidad 0,1

   Encuentre la probabilidad de que dicho comercio pueda vender todos los árboles que tiene almacenados si ha adquirido:

   (a) 1100 árboles

   (b) 1400 árboles

   (c) 1600 árboles

   (d) 2000 árboles

8. La compañía automovilística japonesa Lexus ha ganado una merecida reputación por sus planes de control de calidad. Datos estadísticos recientes muestran que el nuevo modelo Lexus ES 300 presenta:

0 defectos	con probabilidad 0,12
1 defecto	con probabilidad 0,18
2 defectos	con probabilidad 0,25
3 defectos	con probabilidad 0,20
4 defectos	con probabilidad 0,15
5 o más defectos	con probabilidad 0,10

Si usted comprase un nuevo Lexus ES 300, encuentre la probabilidad de que tenga:

(a) 2 o menos defectos

(b) 4 o más defectos

(c) entre 1 y 3 defectos (inclusive)

Sea p la probabilidad de que tenga un número par de defectos. Aunque la información suministrada no permite determinar el valor exacto de p, encuentre una cota

(d) máxima

(e) mínima

que sea consistente con los datos anteriores:

9. Cuando se teclea un manuscrito de cinco páginas, una determinada persona comete:

0 errores	con probabilidad 0,20
1 error	con probabilidad 0,35
2 errores	con probabilidad 0,25
3 errores	con probabilidad 0,15
4 o más errores	con probabilidad 0,05

Si se le da el manuscrito a esa persona, encuentre la probabilidad de que cometa:

(a) 3 o menos errores

(b) 2 o menos errores

(c) 0 errores

10. La tabla siguiente es una versión moderna de las *tablas de vida*, originalmente ideadas por John Graunt en 1662. Muestra las probabilidades de que una persona que ha nacido dentro de un grupo muera en su i-ésima década de vida, variando i desde 1 hasta 10. La

primera década comienza en la fecha de su nacimiento y termina en su décimo cumpleaños, y así sucesivamente.

Tabla de vida

Década	Probabilidad de defunción	Década	Probabilidad de defunción
1	0,062	6	0,124
2	0,012	7	0,215
3	0,024	8	0,271
4	0,033	9	0,168
5	0,063	10	0,028

Por ejemplo, la probabilidad de que una persona fallezca cuando tenga entre 50 y 60 años es 0,124. Encuentre la probabilidad de que una persona:

(a) Fallezca entre los 30 y los 60 años.

(b) No sobrepase los 40 años.

(c) Sobrepase los 80 años.

11. Si está nublado o llueve, se retrasará la excursión familiar planeada para mañana. Las previsiones meteorológicas indican que las probabilidades de que llueva son del 40%, las de que esté nublado son del 50%, y las de que esté nublado y llueva son del 20%. ¿Cuál es la probabilidad de posponer la excursión?

12. En el ejemplo 4.3, ¿qué proporción de clientes no dispone de tarjeta American Express ni de VISA?

13. Se estima que un 30% del total de adultos de Estados Unidos son obesos y que un 3% sufre diabetes. Si un 2% de la población sufre simultáneamente obesidad y diabetes, ¿qué porcentaje de la población padece obesidad o diabetes?

14. Las soldaduras de juntas tubulares pueden tener dos tipos de defectos, que se denominarán A y B. Cada soldadura puede tener el defecto A con probabilidad 0,064, y el defecto B con probabilidad 0,043, y ambos defectos con probabilidad 0,025. Encuentre la proporción de soldaduras que:

    (a) Tienen el defecto A o el defecto B.

    (b) No tienen defecto alguno.

15. Los clientes del departamento de caballeros de un gran almacén compran un traje con probabilidad 0,3, compran una corbata con probabilidad 0,2 y compran un traje y una corbata con probabilidad 0,1. ¿Qué proporción de clientes no compra ni traje ni corbata?

16. Anita tiene un 40% de probabilidad de obtener una calificación de sobresaliente en Estadística, un 60% de conseguir un sobresaliente en Física y un 86% de obtener sobresaliente en Estadística o en Física. Calcule la probabilidad de que:

    (a) No obtenga sobresaliente ni en Estadística ni en Física.

    (b) Obtenga sobresaliente en ambas asignaturas.

17. Este problema utiliza un diagrama de Venn para obtener una demostración formal de la regla de adición. Los sucesos A y B están representados por círculos en el diagrama de Venn.

    En términos de A y B, describa las regiones:

    (a) I

    (b) II

    (c) III

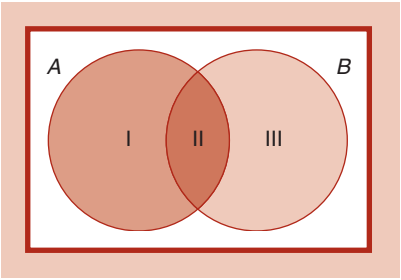


Diagrama de Venn: particionando $A \cup B$.

Exprese, en términos de $P(\text{I})$, $P(\text{II})$ y $P(\text{III})$,

(d) $P(A \cup B)$

(e) $P(A)$

(f) $P(B)$

(g) $P(A \cap B)$

(h) Concluya que

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

4.4 Experimentos con resultados equiprobables

En algunos experimentos es natural asumir que cada resultado posible del espacio muestral tiene la misma probabilidad de ocurrir. Es decir, si el espacio muestral S tiene N resultados.

Digamos $S = \{1, 2, \ldots, N\}$, en ocasiones es razonable suponer que

$$P(\{1\}) = P(\{2\}) = \cdots = P(\{N\})$$

En esta expresión, $P([i])$ es la probabilidad del suceso que contiene únicamente el resultado i; esto es, es la probabilidad de que el resultado del experimento sea i.

Si se usan las propiedades de la probabilidad se puede demostrar que lo anterior implica que la probabilidad de cualquier suceso A es igual a la proporción de resultados del espacio muestral que están en A. Esto es,

$$P(A) = \frac{\text{número de resultados de } S \text{ que están en } A}{N}$$

Ejemplo 4.4 En una muestra de 420 miembros de un centro de jubilados, 144 resultaron ser fumadores; y 276, no fumadores. Si se selecciona a uno de ellos mediante una asignación probabilidades iguales para todos ellos, ¿qué probabilidad hay de que la persona seleccionada sea fumadora?

Solución Existen 420 resultados del espacio muestral del experimento que consiste en la selección. El resultado será la persona seleccionada. Puesto que existen 144 resultados en el suceso consistente en seleccionar un fumador, la probabilidad de este suceso será

$$P\{\text{fumador}\} = \frac{144}{420} = \frac{12}{35}$$ ■

Ejemplo 4.5 Supongamos que, cuando se lanzan dos dados, cada uno de los 36 resultados posibles, dados en el ejemplo 4.1(d), son igualmente probables. Encuentre la probabilidad de que la suma de los dos lanzamientos sea 6 y la probabilidad de que la suma sea 7.

Solución Si A y B denotan los sucesos de que la suma de los dos lanzamientos sea 6 y 7, respectivamente, se tiene que

$$A = \{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\}$$

y

$$B = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$$

Por consiguiente, puesto que A contiene 5 resultados y B contiene 6, se ve que:

$$P(A) = P\{\text{la suma sea 6}\} = 5/36$$

$$P(B) = P\{\text{la suma sea 7}\} = 6/36 = 1/6$$ ■

Ejemplo 4.6 Se desea seleccionar a un hombre y una mujer de un grupo de 10 parejas casadas. Si todas las selecciones posibles son igualmente probables, ¿cuál es la probabilidad de que la mujer y el hombre seleccionados sean pareja?

Solución Una vez que se haya seleccionado al varón, existen 10 posibles selecciones de la mujer. Puesto que sólo una de estas últimas selecciones es la esposa de varón elegido, se ve que la probabilidad que se pide es de 1/10. ■

Cuando todos los resultados del espacio muestral de un experimento son igualmente probables, un elemento seleccionado de ese espacio muestral se dice que ha sido *seleccionado aleatoriamente*.

Ejemplo 4.7 Una escuela elemental ofrece dos asignaturas opcionales de idioma, una de francés y otra de español. Estas asignaturas están abiertas para cualquiera de los 120 estudiantes del último curso. Supongamos que 32 estudiantes se matricularon en la asignatura de francés, 36 en la de español y 8 en ambas. Si se selecciona aleatoriamente a un estudiante entre los 120 antes citados, ¿cuál es la probabilidad de que se haya matriculado en al menos una de estas dos asignaturas?

Solución A y B denotarán los sucesos consistentes en que el estudiante seleccionado se haya matriculado en la clase de francés y en la clase de español, respectivamente. Se determinará $P(A \cup B)$, la probabilidad de que el estudiante esté matriculado en cualquiera de las dos asignaturas de idiomas, usando la regla de adición

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Puesto que 32 de los 120 estudiantes se matricularon en francés, 36 de los 120 se matricularon en español y 8 en ambos idiomas, se tiene que

$$P(A) = \frac{32}{120}, \quad P(B) = \frac{36}{120} \quad \text{y} \quad P(A \cap B) = \frac{8}{120}$$

Por consiguiente,

$$P(A \cup B) = \frac{32}{120} + \frac{36}{120} - \frac{8}{120} = \frac{1}{2}$$

Esto es, la probabilidad de que el estudiante elegido aleatoriamente se haya matriculado en al menos una de las dos asignaturas de idioma es 1/2. ■

Ejemplo 4.8 La tabla 4.3 muestra las frecuencias salariales mayores en 15 años de los trabajadores a tiempo completo, clasificados por salario anual y sexo. Supongamos que se selecciona aleatoriamente a uno de estos trabajadores. Encuentre la probabilidad de que esta persona sea

(a) una mujer

(b) un hombre

Tabla 4.3 Salarios de los trabajadores por sexo, 1989

Clases salariales (en miles de $)	Número		Distribución (porcentaje)	
	Mujeres	Hombres	Mujeres	Hombres
<5	427 000	548 000	1,4	1,1
5–10	440 000	358 000	1,4	,7
10–15	1 274 000	889 000	4,1	1,8
15–20	1 982 000	1 454 000	6,3	2,9
20–30	6 291 000	5 081 000	20,1	10,2
30–40	6 555 000	6 386 000	20,9	12,9
40–50	5 169 000	6 648 000	16,5	13,4
50–100	8 255 000	20 984 000	26,3	42,1
<100	947 000	7 377 000	3,0	14,9
Total	31 340 000	49 678 000	100,0	100,0

Fuente: Departamento de Comercio, Oficina de Censos.

(c) Un hombre con salario inferior a 30 000 $

(d) Una mujer son salario superior a 50 000 $

Solución

(a) Puesto que 31 340 000 de los 31 340 000 + 49 678 000 = 81 018 000 trabajadores son mujeres, la probabilidad de que un trabajador seleccionado aleatoriamente resulte ser una mujer es

$$\frac{31\,340\,000}{81\,018\,000} \approx 0{,}3868$$

Esto es, la probabilidad citada es aproximadamente igual al 38,7%.

(b) Puesto que el suceso consistente en que el trabajador seleccionado aleatoriamente sea un hombre es el complementario del suceso consistente en que el resultado de la selección aleatoria sea una mujer, se desprende de (a) que la probabilidad buscada es aproximadamente $1 - 0{,}3868 = 0{,}6132$.

(c) Como el número de hombres con salario inferior a 30 000 $ es

$$548 + 358 + 889 + 1454 + 5081 = 8330$$

se ve que la probabilidad pedida es de 8330/81 018 ≈ 0,1028. Esto es, existe aproximadamente un 10,3% de probabilidad de que la persona seleccionada sea un hombre con un salario inferior a 30 000 $.

(d) La probabilidad de que la persona seleccionada sea una mujer con un salario superior a 50 000 $ es

$$\frac{8255 + 947}{81\,018} \approx 0{,}1136$$

Esto es, la probabilidad citada es aproximadamente del 11,4%. ■

Problemas

1. En un experimento relacionado con los detectores de humo, se hizo que la alarma sonara en los dormitorios de una residencia universitaria. Entre los 216 residentes, 128 no se despertaron. Si se elige a uno de los residentes aleatoriamente, ¿qué probabilidad hay de que se despertara con el sonido de la alarma?

2. Entre 32 seguidores de una dieta de alimentación con hábitos similares, 18 perdieron peso, 5 ganaron peso y 9 continuaron con el mismo peso. Si se elige a una de esas personas aleatoriamente, encuentre la probabilidad de que:

   (a) Ganara peso.

   (b) Perdiera peso.

   (c) Ni perdiera ni ganara peso.

3. Se extrae una carta aleatoriamente de una baraja ordinaria de 52 cartas. Encuentre la probabilidad de que la carta seleccionada sea:

   (a) un as (b) distinta de un as

   (c) una espada (d) el as de espadas

4. En la tabla siguiente se muestran los 10 países con mayor producción de carne.

País	Producción de carne (en miles de toneladas métricas)
China	20 136
Estados Unidos	17 564
Rusia	12 698
Alemania	6 395
Francia	3 853
Brasil	3 003
Argentina	2 951
Gran Bretaña	2 440
Italia	2 413
Australia	2 373

Supongamos que se ha constituido un Comité de la Organización Mundial de la Salud para analizar las consecuencias que, a largo plazo, se derivan de producir tales cantidades de carne. Dicho comité está compuesto por un representante de cada uno de los países citados. Si se va a elegir al presidente aleatoriamente entre sus miembros, encuentre la probabilidad

de que la presidencia recaiga sobre el miembro de un país cuya producción de carne (en miles de toneladas métricas):

(a) Sea superior a 10 000.

(b) Sea inferior a 3500.

(c) Esté comprendida entre 4000 y 6000.

(d) Esté por debajo de 2000.

5. A continuación se muestran los cinco países que han sido los mayores productores de vehículos a motor en 2002.

Mayores productores de vehículos a motor, 2002

	Total	Coches	Camiones
Estados Unidos	12 328 305	5 027 425	7 300 881
Japón	10 239 949	8 618 725	1 621 224
Alemania	5 469 564	5 122 894	346 700
Francia	3 660 985	3 284 000	376 985
Corea del Sur	3 147 584	2 651 273	496 311

Fuente: Centro de Datos de Noticias y de Sistemas de Comercialización de Automóviles GmbH.

Si se elige aleatoriamente un coche producido en uno de estos países,

(a) ¿Cuál es la probabilidad de que haya sido producido en Estados Unidos?

(b) ¿Cuál es la probabilidad de que provenga de Corea del Sur?

6. Se extrae una moneda de una bolsa que contiene cierto número de monedas de 1 céntimo y cuatro veces más de monedas de 10 céntimos. Si se asume que todas las monedas tienen la misma probabilidad de ser seleccionadas, ¿cuál es la probabilidad de que la moneda extraída sea de 10 céntimos?

7. Un total de 44 de los 100 pacientes de un centro de rehabilitación deben seguir un programa especial que consiste en recibir clases de natación y clases de calistenia. Cada uno de esos 44 pacientes sigue al menos una de estas clases. Supongamos que 26 pacientes asisten a las clases de natación y que 28 pacientes asisten a las clases de calistenia. Encuentre la probabilidad de que un paciente elegido aleatoriamente entre todos los pacientes del centro:

(a) No siga el programa especial de rehabilitación.

(b) Siga simultáneamente las dos clases citadas.

8. Entre las familias de una determinada comunidad, un 20% tiene gatos, un 32% tiene perros y un 12% tiene gatos y perros.

(a) Si se elige aleatoriamente a una familia, ¿cuál es la probabilidad de que no tenga ni gatos ni perros?

(b) Si en la comunidad existen 1000 familias, ¿cuántas de ellas tienen gatos o perros?

9. Entre las estudiantes de una escuela femenina, un 60% no lleva ni anillos ni collares, un 20% lleva anillos y un 30% lleva collares. Si se elige a una estudiante aleatoriamente, encuentre la probabilidad de que lleve:

   (a) anillo o collar

   (b) anillo y collar

10. Un club de deportes tiene 120 miembros: 44 juegan al tenis, 30 juegan al squash y 28 juegan tanto al tenis como al squash. Si se selecciona aleatoriamente a un miembro del club, encuentre la probabilidad de que esta persona:

    (a) No juegue al tenis.

    (b) No juegue al squash.

    (c) No juegue al tenis ni al squash.

11. En el problema 10, ¿cuántos miembros del club juegan al tenis o al squash?

12. Si se lanzan dos dados, encuentre la probabilidad de que la suma de los resultados sea:

    (a) 7 o 11

    (b) uno de los valores 2, 3 o 12

    (c) un número par

13. Suponga que dos personas son seleccionadas aleatoriamente entre un conjunto de 20, que conforman 10 parejas casadas. ¿Cuál es la probabilidad de que las dos personas elegidas estén casadas entre ellas? (*Sugerencia*: Una vez elegida la primera persona, es igualmente probable que la segunda sea una persona cualquiera de las restantes.)

14. En el ejemplo 4.8, encuentre la probabilidad de que un trabajador seleccionado aleatoriamente:

    (a) Gane menos de 15 000 $.

    (b) Sea una mujer con un salario que esté entre 20 000 $ y 40 000 $.

    (c) Gane menos de 50 000 $.

15. Un agente inmobiliario tiene un conjunto de 10 llaves, y una de ellas abre la puerta delantera de la casa que va a enseñar a un cliente. Si las llaves se prueban en un orden completamente aleatorio, encuentre la probabilidad de que:

    (a) La primera llave probada abra la puerta.

    (b) Se prueben las 10 llaves.

16. Un grupo de 5 chicas y 4 chicos se colocan en fila aleatoriamente.

    (a) ¿Cuál es la probabilidad de que la persona colocada en la segunda posición sea un chico?

    (b) ¿Cuál es la probabilidad de que Carlos (uno de los chicos) se encuentre en la segunda posición?

17. Los datos siguientes provienen de Administración Nacional Oceánica y Atmosférica de Estados Unidos. Muestran el número medio de días de cada mes con una precipitación de 0,01 pulgadas o más en Washington, D.C.

Ene.	Feb.	Mar.	Abr.	May.	Jun.	Jul.	Ago.	Sep.	Oct.	Nov.	Dic.
10	9	11	10	11	10	10	9	8	7	8	9

Calcule la probabilidad de que una persona que va a visitar Washington encuentre lluvia el próximo día:

(a) 5 de enero

(b) 12 de agosto

(c) 15 de abril

(d) 15 de mayo

(e) 12 de octubre

4.5 Probabilidad condicionada e independencia

En ocasiones, uno está interesado en calcular probabilidades cuando dispone de cierta información parcial relativa al resultado del experimento. En tales situaciones, las probabilidades se denominan *probabilidades condicionadas*.

Como ejemplo de probabilidad condicionada, supongamos que lanzamos dos dados. Como se indicó en el ejemplo 4.1(d), el espacio muestral de este experimento es el conjunto de los 36 resultados (i, j), donde tanto i como j varían entre 1 y 6. El resultado (i, j) indica que el lanzamiento del primer dado es i y que el del segundo es j.

Supongamos que cada uno de los 36 resultados tiene la misma probabilidad de ocurrir, igual a 1/36. (Cuando esto es así, se dice que el dado está bien construido.) Supongamos, adicionalmente, que se sabe que el primer lanzamiento ha sido un 4. Dada esta información, ¿cuál es la probabilidad de que la suma de los dos lanzamientos sea 10? Para determinar esta probabilidad, se puede razonar como sigue. Dado que el primer lanzamiento ha resultado ser un 4, existen 6 posibles resultados del experimento, que son:

$$(4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6)$$

Además, puesto que estos resultados tienen inicialmente la misma probabilidad de ocurrencia, continuarán teniendo probabilidades iguales. Esto es, dado que el primer lanzamiento ha sido 4, la probabilidad *condicionada* de los resultados de los segundos lanzamientos será 1/6. Puesto que en solamente uno, exactamente el (4, 6), de estos últimos resultados hace que la suma de los dos lanzamientos sea 10, se sigue que la probabilidad de que la suma sea 10, dado que el primer lanzamiento ha sido 4, es 1/6.

Si B denota el suceso consistente en que la suma de los dos lanzamientos sea 10 y A denota el suceso definido por el hecho de que el primer lanzamiento sea 4, la probabilidad obtenida anteriormente se conoce como *probabilidad condicionada de B dado que ha ocurrido A*. Se denota por

$$P(B|A)$$

Una fórmula general para $P(B|A)$ se puede obtener mediante un razonamiento similar al utilizado anteriormente. Supongamos que el resultado del experimento está contenido en A. Para que el resultado esté también en B debe estar simultáneamente en A y en B; esto es, debe estar en $A \cap B$. Sin embargo, como sabemos que el resultado está en A, se tiene que A se convierte en nuestro nuevo (o reducido) espacio muestral, y la probabilidad de que el suceso $A \cap B$ ocurra es la probabilidad de $A \cap B$ relativa a la probabilidad de A. Es decir (véase la figura 4.6):

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Esta definición de la probabilidad condicionada es coherente con la interpretación de la probabilidad como la frecuencia relativa a largo plazo. Para demostrar esto, supongamos que se llevan a cabo un gran número, digamos n, de repeticiones del experimento. Si consideramos solamente aquellos experimentos en los que ocurre el suceso A, $P(B|A)$ será igual a la proporción de ellos en los que también ocurre B. Para ver esto, observe que, puesto que $P(A)$ es la proporción, a largo plazo, de experimentos en los que ocurre A, se tendrá que en n repeticiones del experimento, A ocurrirá aproximadamente $nP(A)$ veces. De igual forma, en aproximadamente $nP(A \cap B)$ de estos experimentos ocurrirán simultáneamente A y B. De aquí se deduce que, entre los aproximadamente $nP(A)$ experimentos cuyos resultados están contenidos en A, aproximadamente $nP(A \cap B)$ de ellos tendrán resultados contenidos también en B. Por consiguiente, de todos aquellos experimentos cuyos resultados están contenidos en A, la proporción de ellos cuyos resultados están también en B es aproximadamente igual a

$$\frac{nP(A \cap B)}{nP(A)} = \frac{P(A \cap B)}{P(A)}$$

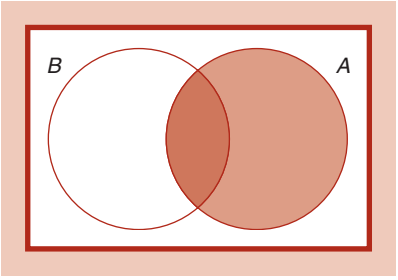


Figura 4.6 $P(B|A) = \frac{P(A \cap B)}{P(A)}$.

Puesto que esta aproximación se hace más exacta a medida que n crece se ve que la definición anterior de probabilidad condicionada de B, dado que A ha ocurrido, es apropiada.

Ejemplo 4.9 Como comprobación adicional de la anterior fórmula de probabilidad condicionada, utilícela para calcular la probabilidad de que la suma de dos lanzamientos de dados sea 10 condicionada a que el primer lanzamiento haya sido 4.

Solución Si B denota el suceso de que la suma de los dos lanzamientos sea 10 y A denota el suceso de que el primer lanzamiento haya sido 4, se tiene que

$$\begin{aligned} P(B|A) &= \frac{P(A \cap B)}{P(A)} \\ &= \frac{P(\{(4, 6)\})}{P(\{(4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6)\})} \\ &= \frac{1/36}{6/36} = \frac{1}{6} \end{aligned}$$

Por consiguiente, se obtiene el mismo resultado que antes. ■

Ejemplo 4.10 La empresa en la que Jacob trabaja está organizando una cena de padres e hijas para aquellos empleados que tienen al menos una hija. La empresa pide a cada uno de sus empleados que asista a la cena con una de sus hijas. Si se sabe que Jacob tiene dos descendientes y que ha sido invitado a la cena. ¿Cuál es la probabilidad de que sus dos descendientes sean hijas? Asuma que el espacio muestral S es

$$S = \{(g, g), (g, b), (b, g), (b, b)\}$$

y que todos los resultados son igualmente probables; donde, por ejemplo, el resultado (g, b) representa que el descendiente mayor de Jacob es una chica y el menor es un chico.

Solución Puesto que Jacob ha sido invitado a la cena, se sabe que al menos uno de sus descendientes es una chica. Si B denota el suceso de que ambos descendientes sean chicas y A el suceso de que al menos uno de sus descendientes sea chica, se ve que la probabilidad pedida es $P(B|A)$. Ésta se puede calcular como sigue:

$$\begin{aligned} P(B|A) &= \frac{P(A \cap B)}{P(A)} \\ &= \frac{P(\{g, g\})}{P(\{(g, g), (g, b), (b, g)\})} \end{aligned}$$

$$= \frac{1/4}{3/4} = \frac{1}{3}$$

Esto es, la probabilidad condicionada de que los dos descendientes de Jacob sean chicas es 1/3, dado que al menos uno de los descendientes es chica. Muchos estudiantes suponen incorrectamente que esta probabilidad condicionada es 1/2, con el razonamiento de que existe la misma probabilidad de que el descendiente que no asiste a la cena sea un chico o una chica. Su equivocación consiste en asumir que ambas posibilidades son igualmente probables, puesto que inicialmente existen 4 resultados igualmente probables. La información de que al menos uno de los descendientes es chica equivale a saber que el resultado no ha sido (*b, b*). Así pues, nos quedan tres resultados igualmente probables, (*g, g*), (*g, b*) y (*b, g*), con lo que se prueba que sólo existe una probabilidad de 1/3 de que Jacob tenga dos hijas. ■

Ejemplo 4.11 La tabla 4.4 muestra el número (en miles) de estudiantes matriculados en una universidad del Estado de California, clasificados por sexo y edad.

(a) Supongamos que se selecciona aleatoriamente a un estudiante. ¿Cuál es la probabilidad de que este estudiante sea una mujer?

Tabla 4.4 Matrículas

Sexo y edad	
Total	12 544
Varones	5 881
de 14 a 17 años	91
de 18 a 19 años	1 309
de 20 a 21 años	1 089
de 22 a 24 años	1 080
de 25 a 29 años	1 016
de 30 a 34 años	613
de 35 años en adelante	684
Mujeres	6 663
de 14 a 17 años	119
de 18 a 19 años	1 455
de 20 a 21 años	1 135
de 22 a 24 años	968
de 25 a 29 años	931
de 30 a 34 años	716
de 35 años en adelante	1 339

Encuentre la probabilidad condicionada de que un estudiante elegido aleatoriamente:

(b) Tenga una edad por encima de los 35 años, dado que este estudiante es un hombre.

(c) Tenga una edad por encima de los 35 años, dado que este estudiante es una mujer.

(d) Sea una mujer, dado que este estudiante tiene más de 35 años.

(e) Sea un hombre, dado que este estudiante tiene entre 20 y 21 años.

Solución

(a) Puesto que hay 6663 mujeres entre el total de estudiantes, que es de 12 544, se sigue que la probabilidad de que el estudiante seleccionado aleatoriamente sea una mujer es

$$\frac{6663}{12\,544} = 0{,}5312$$

(b) Puesto que hay un total de 5881 estudiantes varones, de los que 684 tienen una edad por encima de los 35 años, la probabilidad condicionada pedida es

$$\text{P(tener más de 35 años/ser varón)} = \frac{684}{5881} = 0{,}1163$$

(c) Por un razonamiento análogo al empleado en (b), se ve que

$$\text{P(tener más de 35 años/ser mujer)} = \frac{1339}{6663} = 0{,}2010$$

(d) Puesto que existe un total de 684 + 1339 = 2023 estudiantes con una edad superior a los 35 años, de los que 1339 son mujeres, se tiene que

$$\text{P(ser mujer/tener más de 35 años)} = \frac{1339}{2023} = 0{,}6619$$

(e) Puesto que existe un total de 1089 + 1135 = 2224 estudiantes con una edad comprendida entre los 20 y los 21 años, de los que 1089 son varones, se tiene que

$$\text{P(ser varón/tener una edad entre 20 y 21 años)} = \frac{1089}{2224} = 0{,}4897$$ ■

Dado que

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

se obtiene, tras multiplicar ambos términos por $P(A)$, el siguiente resultado conocido como *regla de multiplicación*.

Regla de multiplicación

$$P(A \cap B) = P(A)P(B|A)$$

Esta regla establece que la probabilidad de que ocurran simultáneamente A y B es igual a la probabilidad de que ocurra A multiplicada por la probabilidad condicionada de B, dado que haya ocurrido A. Esta regla es, a menudo, útil para calcular la probabilidad de una intersección.

Ejemplo 4.12 Supongamos que se elijen aleatoriamente a dos personas de un grupo formado por 4 mujeres y 6 hombres.

(a) ¿Cuál es la probabilidad de que ambas personas sean mujeres?

(b) ¿Cuál es la probabilidad de una sean una mujer y un hombre?

Solución

(a) A y B denotan los sucesos de que la primera persona seleccionada sea una mujer y de que la segunda persona seleccionada sea una mujer, respectivamente. Para calcular la probabilidad pedida $P(A \cap B)$, se comenzará con la identidad

$$P(A \cap B) = P(A)P(B|A)$$

Ahora bien, puesto que las 10 personas, de las que 4 son mujeres, tienen la misma probabilidad de ser el primer individuo seleccionado, se tiene que

$$P(A) = \frac{4}{10}$$

Ahora bien, dado que la primera persona seleccionada es una mujer, se concluye que, en la siguiente selección, las 9 personas restantes, de las cuales 3 son mujeres, son igualmente probables. Por consiguiente,

$$P(B|A) = \frac{3}{9}$$

y, por tanto,

$$P(A \cap B) = \frac{4}{10} \cdot \frac{3}{9} = \frac{2}{15}$$

(b) Para determinar la probabilidad de que la pareja elegida sea de una mujer y un hombre, observe primero que esto puede ocurrir de dos formas disjuntas. La primera persona elegida es un hombre y la segunda una mujer, o bien ocurre en el orden contrario. Si A

denota el suceso consistente en que la primera persona elegida sea un hombre y B denota el suceso de que la segunda persona elegida sea una mujer, se tiene que

$$P(A \cap B) = P(A)P(B|A)$$

Ahora bien, puesto que las 10 personas, de las cuales 6 son hombres, tienen la misma probabilidad de ser elegidas en primer lugar, se tiene que

$$P(A) = \frac{6}{10}$$

Adicionalmente, si la primera persona seleccionada es un hombre, las nueve personas restantes, de las cuales 4 son mujeres, tienen la misma probabilidad de ser seleccionadas a continuación, de donde se deduce que

$$P(B|A) = \frac{4}{9}$$

En consecuencia,

$$P(\text{primero hombre y luego mujer}) = P(A \cap B) = \frac{6}{10} \cdot \frac{4}{9} = \frac{4}{15}$$

Por un razonamiento similar, la probabilidad de que la primera persona elegida sea mujer y la segunda hombre es

$$P(\text{primero mujer y luego hombre}) = \frac{4}{10} \cdot \frac{6}{9} = \frac{4}{15}$$

Dado que el suceso consistente en que la pareja elegida esté compuesta por una mujer y un hombre es la unión de los dos sucesos disjuntos anteriores, se sigue que

$$P(\text{1 mujer y 1 hombre}) = \frac{4}{15} + \frac{4}{15} = \frac{8}{15}$$ ■

Por lo general, la probabilidad condicionada de que ocurra B dado que haya ocurrido A no tiene por qué coincidir con la probabilidad (incondicional) de B. Es decir, saber que ha ocurrido A generalmente hace cambiar la probabilidad de ocurrencia de B. Cuando $P(B/A)$ es igual a $P(B)$, se dice que B es *independiente* de A.

Puesto que

$$P(A \cap B) = P(A)P(B|A)$$

se deduce que B es independiente de A si

$$P(A \cap B) = P(A)P(B)$$

Ya que esta última ecuación es simétrica en A y B, se tiene que, si B es independiente de A, también A es independiente de B.

También se puede demostrar que si A y B son independientes, la probabilidad de B dado que A no ocurra es igual a la probabilidad (incondicional) de B. Esto es, si A y B son independientes, se cumple que

$$P(B|A^c) = P(B)$$

Así pues, cuando A y B son independientes, cualquier información acerca de la ocurrencia o no ocurrencia de uno de estos sucesos no afecta a la probabilidad del otro.

Los sucesos A y B son *independientes* si

$$P(A \cap B) = P(A)P(B)$$

Si A y B son independientes, la probabilidad de que uno de ellos ocurra no se ve afectada por la información de que el otro haya ocurrido o no.

Ejemplo 4.13 Supongamos que se lanzan dos dados bien construidos, de modo que cada uno de los 36 resultados posibles son igualmente probables. A denotará el suceso consistente en que resulte un 3 en el primer lanzamiento, B denotará el suceso consistente en que la suma de los dos lanzamientos sea 8, y C denotará el hecho de que la suma de los dos lanzamientos sea 7.

(a) ¿ A y B son independientes?

(b) ¿A y C son independientes?

Solución

(a) Puesto que $A \cap B$ es el suceso de que se obtenga un 3 en el primer lanzamiento y un 5 en el segundo, se ve que

$$P(A \cap B) = P(\{(3,5)\}) = \frac{1}{36}$$

Por otra parte,

$$P(A) = P(\{(3,1),(3,2),(3,3),(3,4),(3,5),(3,6)\}) = \frac{6}{36}$$

y

$$P(B) = P(\{(2,6),(3,5),(4,4),(5,3),(6,2)\}) = \frac{5}{36}$$

Por consiguiente, puesto que $1/36 \neq (6/36) \cdot (5/36)$, se ve que

$$P(A \cap B) \neq P(A)P(B)$$

de donde se deduce que A y B no son independientes.

Intuitivamente, la razón de que los sucesos no sean independientes es que la probabilidad de que la suma de los dos lanzamientos sea 8 se ve afectada por el resultado del primer lanzamiento. En particular, la probabilidad de que la suma sea 8 se ve incrementada por el conocimiento de que se ha obtenido un 3 en el primer lanzamiento, puesto que aún así es posible que la suma de los dos lanzamientos sea 8 (lo cual no sería posible si se hubiera obtenido un 1 en el primer lanzamiento).

(b) Los sucesos A y C son independientes. Basta con observar que

$$P(A \cap C) = P(\{3, 4\}) = \frac{1}{36}$$

mientras que

$$P(A) = \frac{1}{6}$$

y

$$P(C) = P(\{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}) = \frac{6}{36}$$

En consecuencia,

$$P(A \cap C) = P(A)P(C)$$

de donde se deduce que A y C son independientes.

Es bastante intuitivo razonar que el suceso obtener un 7 como suma de los dos lanzamientos es independiente del suceso obtener un 3 en el primer lanzamiento. Basta observar que no importa qué resultado se obtenga en el primer lanzamiento, siempre existe un único resultado del segundo lanzamiento con el que se consigue que la suma sea 7. Como resultado, la probabilidad condicionada de que la suma sea 7 dado el valor del primer lanzamiento es siempre igual a 1/6. ■

Ejemplo 4.14 Considere la tabla 4.4, presentada en el ejercicio 4.11. Supongamos que se eligen aleatoriamente a un estudiante mujer e, independientemente, a un estudiante varón. Calcule la probabilidad de que ambos estudiantes tengan una edad comprendida entre los 22 y los 24 años.

Solución Dado que 1080 de los 5881 estudiantes varones tienen una edad comprendida entre los 22 y los 24 años, se sigue que

$$\text{P}(\{\text{elegir un varón con edad entre 22 y 24 años}\}) = \frac{1080}{5881} \approx 0{,}1836$$

De igual forma, puesto que 968 de las 6663 estudiantes mujeres tienen una edad comprendida entre los 22 y los 24 años, se ve que

$$\text{P}(\{\text{elegir una mujer con edad entre 22 y 24 años}\}) = \frac{968}{6663} \approx 0{,}1453$$

Sin olvidar que los estudiantes se han extraído de forma independiente, se obtiene

$$\text{P}(\{\text{los dos estudiantes elegidos tengan entre 22 y 24 años de edad}\}) = \frac{1080}{5881} \cdot \frac{968}{6663} \approx 0{,}0267$$

Esto es, la probabilidad de que los dos estudiantes seleccionados tengan edades comprendidas entre los 22 y los 24 años es aproximadamente igual al 2,7%. ■

Aunque hasta ahora sólo se ha considerado la independencia de un par de sucesos, este concepto se puede extender a cualquier número de sucesos. La probabilidad de la intersección de cualquier número de sucesos independientes será igual al producto de sus probabilidades.

Si $A_1, \ldots, A_n$ son independientes, se cumple que

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)P(A_2) \cdots P(A_n)$$

Ejemplo 4.15 Una pareja se plantea tener tres hijos. Si se asume que para cada hijo existe la misma probabilidad de que sea varón o sea hembra y que los sexos de los distintos hijos son independientes, encuentre la probabilidad de que:

(a) Los tres hijos sean hembras.

(b) Al menos uno de los hijos sea hembra.

Solución

(a) Si A_i es el suceso consistente en que el i-ésimo hijo sea una hembra, se tendrá

$$\begin{aligned} \text{P(todos hembras)} &= P(A_1 \cap A_2 \cap A_3) \\ &= P(A_1)P(A_2)P(A_3) \text{ por la independencia} \\ &= \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8} \end{aligned}$$

(b) La forma más sencilla de calcular la probabilidad de tener al menos una hija consiste en calcular primero la probabilidad del suceso complementario: que todos los hijos sean varones. Puesto que, por un razonamiento similar al empleado en el apartado (a),

$$P(\text{todos chicos}) = \frac{1}{8}$$

se ve que

$$P(\text{al menos una chica}) = 1 - P(\text{todos chicos}) = \frac{7}{8}$$ ■

Problemas

1. Se estima que un 30% de los adultos de Estados Unidos están obesos, un 3% de ellos padecen diabetes y un 2% simultáneamente son obesos y sufren diabetes. Determine la probabilidad condicionada de que un individuo elegido aleatoriamente:

   (a) Padezca diabetes, dado que es obeso.

   (b) Sea obeso, dado que padece diabetes.

2. Supongamos que una moneda se lanza dos veces, y asumamos que los cuatro resultados posibles son igualmente probables. Si en el primer lanzamiento ha salido cara, encuentre la probabilidad condicionada de que salga cara en los dos lanzamientos.

3. Considere la tabla 4.3 presentada en el ejemplo 4.8. Supongamos que se elige aleatoriamente a uno de los trabajadores. Calcule la probabilidad condicionada de que ese trabajador:

   (a) Sea una mujer, dado que tiene un salario superior a 25 000 $.

   (b) Gane más de 25 000 $, dado que el trabajador es una mujer.

4. Un 52% de los estudiantes de una universidad son mujeres. Un 5% de dichos estudiantes pretenden especializarse en Ciencias de la Computación. Un 2% de los estudiantes son mujeres que se están especializando en Ciencias de la Computación. Si se selecciona aleatoriamente a un estudiante, determine la probabilidad condicionada de que:

   (a) Ese estudiante sea mujer, dado que se está especializando en Ciencias de la Computación.

   (b) Ese estudiante se esté especializando en Ciencias de la Computación, dado que el estudiante es mujer.

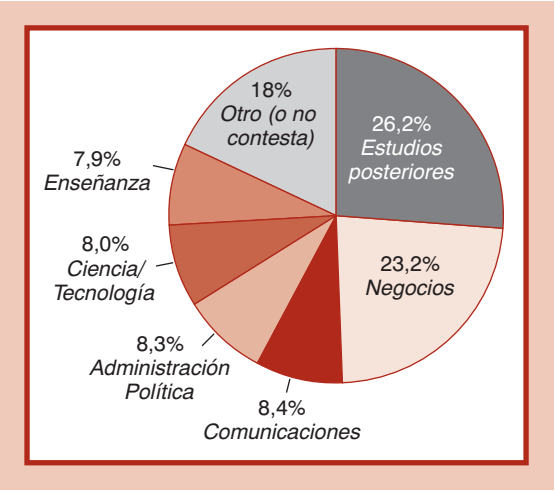
Los problemas 5 y 6 se refieren a los datos de la tabla siguiente, que describe la distribución de edad de los residentes de un condado del norte de California.

Edad	Número
0–9	4200
10–19	5100
20–29	6200
30–39	4400
40–49	3600
50–59	2500
60–69	1800
70 o más	1100

5. Si se elige aleatoriamente a un residente de ese condado, determine la probabilidad de que el residente tenga una edad:
   (a) menor de 10 años
   (b) comprendida entre los 10 y los 20 años
   (c) comprendida entre los 20 y los 30 años
   (d) comprendida entre los 30 y los 40 años
6. Encuentre la probabilidad condicionada de que un residente elegido aleatoriamente tenga una edad:
   (a) Comprendida entre los 10 y los 20 años, dado que el residente tiene menos de 30 años.
   (b) Comprendida entre los 30 y los 40 años, dado que el residente tiene más de 30 años.
7. Un club de juegos de mesa tiene 120 miembros: 40 juegan al ajedrez; 56 juegan al bridge, y 26 juegan tanto al ajedrez como al bridge. Si se elige a un miembro del club aleatoriamente, encuentre la probabilidad condicionada de que:
   (a) Juegue al ajedrez, dado que también juega al bridge.
   (b) Juegue al bridge, dado que también juega al ajedrez.
8. Considere la tabla 4.4 presentada en el ejemplo 4.11. Determine la probabilidad condicionada de que un estudiante elegido aleatoriamente:
   (a) Tenga menos de 25 años, dado que el estudiante es varón.
   (b) Sea un varón, dado que el estudiante tiene menos de 25 años.
   (c) Tenga menos de 25 años, dado que el estudiante es mujer.
   (d) Sea una mujer, dado que el estudiante tiene menos de 25 años.

9. El siguiente gráfico de tarta detalla los planes de futuro de los graduados de la Universidad de Harvard en el año 2004.

Planes de los graduados en 2004 para el año siguiente.



Supongamos que se elige aleatoriamente a un graduado. Dado que ese graduado no se dedicará a los negocios ni a la enseñanza, calcule la probabilidad de que:

(a) Tenga la intención de realizar estudios posteriores.

(b) Planee dedicarse a la enseñanza o bien a los estudios posteriores.

(c) Pretenda dedicarse a las comunicaciones o bien a los estudios posteriores.

(d) No tenga intención de dedicarse a ciencia/tecnología.

(e) No intente dedicarse a las comunicaciones ni a los negocios.

(f) No tenga planeado dedicarse a ciencia/tecnología o a administración/política.

10. Muchos psicólogos creen que el orden de nacimiento y la personalidad están relacionados. Para analizar esta hipótesis se ha seleccionado aleatoriamente a 400 alumnos de una escuela elemental, y se les ha sometido a un test de medición de confianza. Según los resultados del test, cada estudiante fue clasificado como seguro de sí mismo o inseguro, tanto para los que eran primogénitos como para los que no lo eran. La cantidad de estudiantes que están dentro de cada una de las clases posibles se muestra a continuación:

	Primogénito	**No primogénito**
Seguro de sí mismo	62	60
Inseguro	105	173

Por ejemplo, entre los 167 primogénitos, 62 fueron catalogados como seguros de sí mismos. Si suponemos que se selecciona aleatoriamente a un alumno del grupo:

(a) ¿Cuál es la probabilidad de que sea un primogénito?

(b) ¿Cuál es la probabilidad de que el estudiante haya sido catalogado como seguro de sí mismo?

(c) ¿Cuál es la probabilidad condicionada de que sea un estudiante seguro de sí mismo, dado que el estudiante es un primogénito?

(d) ¿Cuál es la probabilidad condicionada de que sea un estudiante seguro de sí mismo, dado que el estudiante no es un primogénito?

(e) ¿Cuál es la probabilidad condicionada de que el estudiante sea un primogénito, dado que ha sido catalogado como seguro de sí mismo?

11. Se eligen aleatoriamente dos cartas de una baraja de 52. ¿Cuál es la probabilidad de que ambas sean ases, dado que son de diferente palo?

12. En las votaciones presidenciales de 1984 en Estados Unidos, un 68,3% de los electores potenciales se registraron como votantes; de estos últimos, realmente votaron un 59,9%. Supongamos que se selecciona aleatoriamente a un elector potencial.

(a) ¿Cuál es la probabilidad de que el elector seleccionado votara realmente?

(b) Si el elector seleccionado no votó realmente, ¿cuál es la probabilidad condicionada de que éste se hubiera registrado como votante?

Observación: Para que los electores potenciales puedan votar previamente se deben registrar como votantes.

13. Hay 30 psiquiatras y 24 psicólogos participando en una conferencia. Se elige aleatoriamente a dos de esas 54 personas para que participen en una mesa de debate. ¿Cuál es la probabilidad de que al menos un psicólogo sea elegido. (*Sugerencia*: Es mejor calcular primero la probabilidad del suceso complementario; es decir, la probabilidad de que no sea elegido ningún psicólogo para la mesa.)

14. Un chico tiene 12 calcetines en un cajón: 5 rojos, 4 azules y 3 verdes. Si se eligen aleatoriamente 2 calcetines, encuentre la probabilidad de que:

(a) Ambos sean rojos.

(b) Ambos sean azules.

(c) Ambos sean verdes.

(d) Los dos sean del mismo color.

15. Se eligen aleatoriamente dos cartas de una baraja de 52 naipes. Encuentre la probabilidad de que:

(a) Ninguna sea de espadas.

(b) Al menos una sea de espadas.

(c) Las dos sean de espadas.

16. Hay n calcetines en un cajón, de los que 3 son rojos. Supongamos que, si se eligen dos calcetines aleatoriamente, la probabilidad de que ambos sean rojos es 1/2. Encuentre n.

*17. Supongamos que la ocurrencia de A hace que B tenga más probabilidad de ocurrir. Demuestre, en este caso, que la ocurrencia de B hace que A tenga mayor probabilidad de ocurrir.

Es decir, demuestre que si

$$P(B|A) > P(B)$$

también es cierto que

$$P(A|B) > P(A)$$

18. Se lanzan dos dados bien construidos.

    (a) ¿Cuál es la probabilidad de que en al menos uno de los lanzamientos salga un 6?

    (b) Si la suma de los resultados de los dos lanzamientos ha sido 9, ¿cuál es la probabilidad condicionada de que se haya obtenido un 6 en al menos uno de los lanzamientos?

    (c) Si la suma de los dos lanzamientos es 10, ¿cuál es la probabilidad condicionada de que en al menos uno de los lanzamientos haya salido un 6?

19. Existe un 40% de posibilidades de que una determinada compañía abra una nueva sucursal en Chicago. Si lo hace, Norris tiene una probabilidad del 60% de ser nombrado director. ¿Cuál es la probabilidad de que Norris sea nombrado director de la sucursal de Chicago?

20. Según un geólogo, la probabilidad de que haya petróleo en una determinada parcela de terreno es 0,7. Adicionalmente, si el petróleo existe, la probabilidad de que se encuentre con la primera perforación es 0,5. ¿Cuál es la probabilidad de que se encuentre petróleo en la primera perforación?

21. En un hospital, la probabilidad de que un paciente fallezca en la mesa de operaciones durante una intervención a corazón abierto es 0,20. Un paciente que sobrevive a la operación tiene un 15% de probabilidades de fallecer en el hospital debido a problemas postoperatorios. ¿Qué porcentaje de pacientes de operaciones a corazón abierto sobrevive tanto a la operación como a los problemas postoperatorios?

22. Una urna contiene inicialmente 4 bolas blancas y 6 negras. Cada vez que se extrae una bola se apunta el color y se devuelve a la urna junto con otra bola del mismo color. ¿Cuál es la probabilidad de que en las dos primeras extracciones se obtengan bolas negras?

23. Reconsidere el problema 7.

    (a) Si se selecciona a un miembro aleatoriamente, ¿cuál es la probabilidad de que la persona elegida juegue al ajedrez o al bridge?

    (b) ¿Cuántos miembros no juegan al ajedrez ni al bridge?

    Si se selecciona a dos miembros aleatoriamente, encuentre la probabilidad de que:

    (c) Ambos jueguen al ajedrez.

    (d) Ninguno de ellos juegue al ajedrez ni al bridge.

    (e) Ambos jueguen al ajedrez o al bridge.

24. Considere la tabla 4.4 del ejemplo 4.11. Supongamos que en 1985 se seleccionó aleatoriamente a dos estudiantes, una mujer y un varón.

(a) Encuentre la probabilidad de que exactamente uno de ellos tuviera más de 30 años.

(b) Dado que exactamente uno de ellos tiene más de 30 años, encuentre la probabilidad condicionada de que el mayor sea el varón.

25. José y Jim van juntos a cazar patos. Supongamos que José tiene una probabilidad 0,3 de dar en el blanco y que Jim, independientemente, tiene una probabilidad 0,1. Los dos han disparado al mismo pato.

(a) Dado que solamente uno de ellos ha acertado, ¿cuál es la probabilidad condicionada de que haya acertado José? ¿Y la de que haya acertado Jim?

(b) Dado que el pato ha sido alcanzado, ¿cuál es la probabilidad condicionada de que José haya acertado? ¿Y la de que haya acertado Jim?

26. Una pareja tiene dos hijos. A denota el suceso consistente en que el hijo mayor sea hembra, y B denota que el hijo menor sea varón. Si se asume que los cuatro posibles resultados son equiprobables, demuestre que los sucesos A y B son independientes.

27. Un modelo simplificado del movimiento de precios de una acción asume que cada día el precio de la acción aumenta 1 unidad con una probabilidad p o que disminuye 1 unidad con una probabilidad $1-p$. Las variaciones del precio en días diferentes se suponen independientes. Finalmente, asumamos que p es igual a 1/2 para una determinada acción. (Por consiguiente, si por ejemplo el precio de la acción al final del día de hoy es de 100 unidades, al final de mañana será de 101 o de 99, con probabilidades iguales.)

(a) ¿Cuál es la probabilidad de que el precio, al cabo de dos días, sea igual al precio original?

(b) ¿Cuál es la probabilidad de que, al cabo de 3 días, el precio de la acción se haya incrementado en 1 unidad?

(c) Si al cabo de 3 días el precio de la acción ha crecido 1 unidad, ¿cuál es la probabilidad condicionada de que el precio haya aumentado el primer día?

28. Se selecciona aleatoriamente a un residente varón de Nueva York. ¿Cuáles de los siguientes pares de sucesos A y B parece razonable que sean independientes?

(a) A: Es un periodista.
B: Tiene ojos marrones.

(b) A: Ayer tuvo dolor de cabeza.
B: Ayer tuvo un accidente.

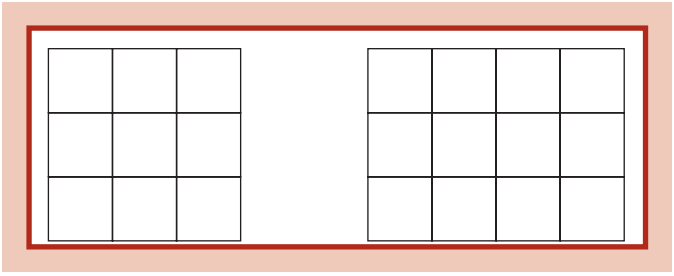
(c) A: Lleva una camisa blanca.
B: Ha llegado tarde al trabajo.

29. Se lanza sucesivamente una moneda, con igual probabilidad de obtener cara y cruz, hasta que se consigue una cruz. Si suponemos que los sucesivos lanzamientos son independientes, ¿cuál es la probabilidad de que la moneda se tenga que lanzar 5 veces como mínimo? (*Sugerencia*: Rellene la palabra que falta en la frase siguiente. La moneda tendrá que lanzarse al menos 5 veces si en los primeros _____ lanzamientos se obtienen caras.)

30. Se lanza un dado hasta que se obtiene un 5. Si asumimos que todas las caras del dado son equiprobables y que los sucesivos lanzamientos son independientes, ¿cuál es la probabilidad de que sea necesario hacer más de 6 lanzamientos?

31. Supongamos que la probabilidad de que un amigo esté comunicando cuando se le intenta llamar por teléfono es de 0,1. ¿Es razonable suponer que la probabilidad de que esté comunicando sea de 0,01 cuando se le llama dos veces, una inmediatamente después de la otra? Si cree que no lo es, ¿se puede imaginar una condición bajo la cual la anterior sea una suposición razonable?

32. Dos superficies contienen 9 y 12 parcelas como se indica a continuación.



Para un experimento agrícola, se selecciona aleatoria e independientemente una parcela de cada superficie.

(a) ¿Cuál es la probabilidad de que las dos parcelas elegidas hagan esquina?

(b) ¿Cuál es la probabilidad de que ninguna de las dos parcelas haga esquina?

(c) ¿Cuál es la probabilidad de que al menos una de las dos parcelas haga esquina?

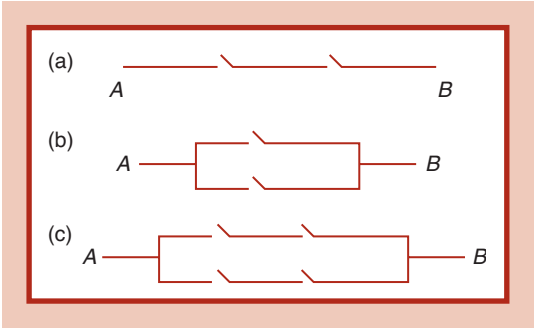
33. Se elige aleatoriamente una carta de una baraja de 52. A es el suceso consistente en que la carta seleccionada sea un as, y B es el de que la carta citada sea de espadas. Demuestre que A y B son independientes.

34. Se lanzan un par de dados bien construidos. A es el suceso de que la suma de los dos dados sea 7. ¿Es A independiente de que el primer dado resulte un 1?, ¿y un 2?, ¿y un 4?, ¿y un 5?, ¿y un 6?

35. ¿Cuál es la probabilidad de que dos desconocidos hayan nacido el mismo día del año?

36. En una publicación de Estados Unidos se indicaba que un 4,78% de los fallecimientos de 1988 se debían a accidentes. ¿Cuál es la probabilidad de que 3 fallecimientos seleccionados aleatoriamente sean todos debidos a accidentes?

37. Cada interruptor de los circuitos siguientes se cierra con probabilidad 0,8. Si todos los interruptores funcionan independientemente, ¿cuál es la probabilidad de que circule la corriente entre A y B en cada circuito? (El circuito de la parte (a) de la figura, que precisa que sus dos relés estén cerrados para que circule la corriente, se denomina *circuito en serie*. El circuito de la parte (b), que necesita que al menos uno de los relés esté cerrado para que pase la corriente, se denomina *circuito en paralelo*.)



(*Sugerencia*: Para las partes (b) y (c), use la regla de adición.)

38. Una urna contiene 5 bolas blancas y 5 negras. Se seleccionan aleatoriamente dos bolas de esta urna. A es el suceso correspondiente a que la primera bola sea blanca y B es el suceso de que la segunda bola sea negra. ¿Los sucesos A y B son independientes? Explique su razonamiento.

39. Suponga, en el problema 38, que la primera bola se remplaza a la urna antes de que se extraiga la segunda. ¿Los sucesos A y B, en este caso, son independientes? Explique, de nuevo, su razonamiento.

40. Suponga que cualquier persona a la que se la pregunta si está a favor de cierta propuesta contesta *sí* con una probabilidad de 0,7, y que responde *no* con una probabilidad de 0,3. Asuma, además, que las respuestas dadas por las distintas personas son independientes. Encuentre la probabilidad de que las siguientes cuatro personas encuestadas:

    (a) Todas den la misma contestación.

    (b) Las dos primeras contesten *no* y las dos últimas respondan *sí*.

    (c) Al menos una conteste *no*.

    (d) Exactamente tres contesten *sí*.

    (e) Al menos una conteste *sí*.

41. Los datos siguientes, provenientes de la Administración Nacional Oceánica y Atmosférica de Estados Unidos, muestran el número medio de días con al menos 0,01 pulgadas de lluvia en varios meses para las ciudades de Mobile, Phoenix y Los Ángeles.

    Número medio de días con 0,01 pulgadas de precipitación o más.

Ciudad	Enero	Abril	Julio
Mobile	11	7	16
Phoenix	4	2	4
Los Ángeles	6	3	1

Supongamos que una persona planea visitar Phoenix el 4 de enero del año próximo, Los Ángeles el 10 de abril y Mobile el 15 de julio.

(a) ¿Cuál es la probabilidad de que llueva en los tres viajes?

(b) ¿Cuál es la probabilidad de que no llueva en los tres viajes?

(c) ¿Cuál es la probabilidad de que llueva en los viajes a Phoenix y Mobile, pero no en el viaje a Los Ángeles?

(d) ¿Cuál es la probabilidad de que llueva en los viajes a Mobile y a Los Ángeles, pero no en el viaje a Phoenix?

(e) ¿Cuál es la probabilidad de que llueva en los viajes a Phoenix y a Los Ángeles, pero no en el viaje a Mobile?

(f) ¿Cuál es la probabilidad de que llueva exactamente en dos de los tres viajes?

42. Cada chip de ordenador producido por la máquina A es defectuoso con probabilidad 0,10; mientras que los chips producidos por la máquina B tienen probabilidad 0,05 de ser defectuosos. Si se toman un chip producido por la máquina A y otro producido por la máquina B, encuentre la probabilidad (asumiendo independencia) de que:

(a) Los dos chips sean defectuosos.

(b) Ninguno de los dos sea defectuoso.

(c) Sólo uno de ellos sea defectuoso.

Si ocurre que sólo uno de los dos chips resulta defectuoso, encuentre la probabilidad de que el chip defectuoso haya sido producido por:

(d) la máquina A

(e) la máquina B

43. Un test genético ha permitido que los padres sepan si sus hijos tienen el riesgo de padecer fibrosis quística (FQ), una enfermedad neuronal degenerativa. Un hijo que reciba un gen FQ de cada progenitor desarrollará la enfermedad en su pubertad y no llegará a la edad adulta. Un hijo que no reciba un gen FQ o sólo reciba un gen FQ no desarrollará la enfermedad; sin embargo, si recibe un solo gen FQ puede transmitírselo a sus descendientes. Si un individuo es portador del gen FQ, cada uno de sus hijos recibirá el gen con una probabilidad de 1/2.

(a) Si los dos padres son portadores del gen FQ, ¿cuál es la probabilidad de que sus hijos desarrollen fibrosis quística?

(b) ¿Cuál es la probabilidad de que una persona de 25 años de edad no sea portadora del gen FQ pero que su hermano sí que lo sea?

*4.6 Teorema de Bayes

Para los sucesos A y B, se verifica que

$$A = (A \cap B) \cup (A \cap B^c)$$

Se puede comprobar que la igualdad anterior es cierta con sólo observar que para que un resultado esté en A debe estar en A y en B o bien debe estar en A pero no en B (véase la

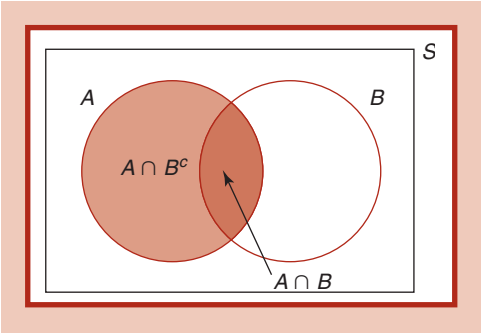


Figura 4.7 $A = (A \cap B) \cup (A \cap B^c)$.

figura 4.7). Puesto que $A \cap B$ y $A \cap B^c$ son mutuamente excluyentes (¿por qué?) se tiene por la propiedad 3 (véase la sección 4.3) que

$$P(A) = P(A \cap B) + P(A \cap B^c)$$

Puesto que

$$P(A \cap B) = P(A|B)P(B) \qquad \text{y} \qquad P(A \cap B^c) = P(A|B^c)P(B^c)$$

se ha demostrado la siguiente igualdad:

$$P(A) = P(A|B)P(B) + P(A|B^c)\,P(B^c) \tag{4.1}$$

Esta igualdad establece que la probabilidad de un suceso A es igual a la media ponderada de las probabilidades condicionadas de que ocurra A dado que B haya ocurrido y de que ocurra A dado B no haya ocurrido; cada una de estas probabilidades condicionadas tiene un peso igual a la probabilidad del suceso condicionante. Esta es una fórmula muy útil porque nos permite calcular la probabilidad de cualquier suceso A "condicionando" primero por los hechos de que otro suceso cualquiera B haya ocurrido o no.

Antes de ilustrar la utilidad de la ecuación (4.1) se considerará el problema de cómo reevaluar una probabilidad inicial a la luz de una evidencia adicional. Supongamos que se está estudiando una cierta hipótesis; supongamos que H denota el suceso de que la hipótesis es cierta y que $P(H)$ denota la probabilidad de que sea cierta. Ahora, supongamos que se dispone de una evidencia adicional, llamémosla E, concerniente a la hipótesis citada. En consecuencia, se desearía determinar $P(H|E)$, la probabilidad condicionada de que la hipótesis es cierta, dada la evidencia adicional E. Se tiene, por la definición de la probabilidad condicionada,

$$P(H|E) = \frac{P(H \cap E)}{P(E)}$$

$$= \frac{P(E|H)P(H)}{P(E)}$$

Si se usa la ecuación 4.1, se puede calcular $P(E)$ condicionando por los hechos de que la hipótesis sea cierta y no sea cierta. Esto conduce a la siguiente identidad, conocida como *teorema de Bayes.*

Teorema de Bayes

$$P(H|E) = \frac{P(E|H)P(H)}{P(E|H)P(H) + P(E|H^c)P(H^c)}$$

Ejemplo 4.16 Una compañía de seguros cree que la gente puede ser clasificada en dos grupos: aquellos que tienen predisposición a tener accidentes y aquellos que no la tienen. Los datos indican que una persona predispuesta a tener accidentes sufrirá un accidente en el plazo de 1 año con probabilidad 0,1; esta misma probabilidad para las personas de la otra clase es 0,05. Supongamos que la probabilidad de que un nuevo asegurado sea propenso a tener accidentes es 0,2.

(a) ¿Cuál es la probabilidad de que un nuevo asegurado sufra un accidente en el plazo de un año?

(b) Si un nuevo asegurado tiene un accidente en el primer año, ¿cuál es la probabilidad de que sea propenso a tener accidentes?

Solución H es el suceso de que un nuevo asegurado sea propenso a los accidentes, y A es el suceso de que tenga un accidente durante el primer año. Se puede calcular $P(A)$ condicionando sobre si la persona es o no propensa a tener accidentes:

$$\begin{aligned} P(A) &= P(A|H)P(H) + P(A|H^c)P(H^c) \\ &= (0,1)(0,2) + (0,05)(0,8) = 0,06 \end{aligned}$$

Por consiguiente, existe una probabilidad del 6% de que un nuevo asegurado tenga un accidente durante el primer año.

La probabilidad $P(H/A)$ se puede calcular como sigue:

$$\begin{aligned} P(H|A) &= \frac{P(H \cap A)}{P(A)} \\ &= \frac{P(A|H)P(H)}{P(A)} \\ &= \frac{(0,1)(0,2)}{0,06} = \frac{1}{3} \end{aligned}$$

Así pues, si un nuevo asegurado tiene un accidente durante el primer año, la probabilidad condicionada de que el asegurado sea propenso a tener accidentes es 1/3. ■

Ejemplo 4.17 Si cierta enfermedad está presente, existe una probabilidad del 99% de que una prueba sanguínea sea efectiva para detectarla. Sin embargo, la prueba también ofrece un resultado *positivo falso* en un 2% de los pacientes sanos. (Es decir, si una persona sana se

somete a la prueba, existe una probabilidad igual a 0,02 de que la prueba indique que esa persona está enferma.) Supongamos que un 0,5% de la población padece la enfermedad. Encuentre la probabilidad condicionada de que un individuo que aleatoriamente se ha sometido a la prueba padezca realmente la enfermedad, dado que su prueba ha resultado positiva.

Solución D denota el suceso de que la persona padezca la enfermedad, y E denota que la prueba resulta positiva. Se pretende determinar $P(D/E)$, lo cual se puede conseguir si se aplica el teorema de Bayes como sigue:

$$P(D|E) = \frac{P(E|D)P(D)}{P(E|D)P(D) + P(E|D^c)P(D^c)}$$

$$= \frac{(0{,}99)(0{,}005)}{(0{,}99)(0{,}005) + (0{,}02)(0{,}995)} = 0{,}199$$

Así pues, existe una probabilidad aproximada del 20% de que una persona de la población, aleatoriamente elegida, cuya prueba haya resultado positiva padezca realmente la enfermedad. (La razón por la que esta probabilidad es tan baja se debe a que la probabilidad de que la persona elegida no padezca la enfermedad, aun cuando dé positivo en la prueba, es mayor que la probabilidad de que la persona padezca la enfermedad y dé positivo en la prueba.) ■

Problemas

1. Hay dos monedas sobre una mesa. Cuando se lanzan, la probabilidad de que salga cara es 0,5 para una moneda, mientras que para la otra es 0,6. Se selecciona aleatoriamente una de las monedas y se lanza.
   (a) ¿Cuál es la probabilidad de que salga cara?
   (b) Si sale cruz, ¿cuál es la probabilidad de que la moneda lanzada sea la que está bien construida? (Es decir, aquélla cuyos resultados cara y cruz son igualmente probables.)
2. Supongamos que, cuando un estudiante responde a una pregunta de un test de elección múltiple, o bien conoce la respuesta o bien la contesta al azar. Si contesta al azar, la probabilidad de que sea la respuesta correcta es 1/5. Si la probabilidad de que un estudiante conozca la contestación es 0,6, ¿cuál es la probabilidad condicionada de que el estudiante conozca la contestación, dado que ha contestado correctamente?
3. Una inspectora a cargo de una investigación criminal tiene una certeza del 60% de la culpabilidad de un sospechoso. Se acaba de descubrir un hecho que evidencia que el criminal es zurdo. Aunque la inspectora sabe que un 18% de las personas son zurdas, le gustaría saber si el sospechoso es zurdo.
   (a) ¿Cuál es la probabilidad de que el sospechoso sea zurdo?
   (b) Si el sospechoso resulta ser zurdo, ¿cuál es la probabilidad de que el sospechoso sea culpable?

4. La urna 1 contiene 4 bolas rojas y 3 azules, y la urna 2 contiene 2 bolas rojas y 2 azules. Se selecciona aleatoriamente una bola de la urna 1 y se coloca en la urna 2. Después se extrae una bola de la urna 2.

   (a) ¿Cuál es la probabilidad de que la bola extraída de la urna 2 sea roja?

   (b) Si la bola extraída de la urna 2 ha sido azul, ¿cuál es la probabilidad condicionada de que la bola extraída de la urna 1 sea roja?

5. Considere un test de diagnóstico cuya seguridad es del 97%, tanto para los que padecen la enfermedad como para los que no la padecen. (Esto es, si una persona padece la enfermedad, el diagnóstico es positivo con probabilidad 0,97; y, si la persona no la padece, el diagnóstico es negativo igualmente con probabilidad 0,97.) Supongamos que un 2% de la población sufre la enfermedad. Si el diagnóstico resulta positivo para una persona de la población seleccionada aleatoriamente, ¿cuál es la probabilidad condicionada de que esa persona padezca la enfermedad?

6. Se tienen tres cartas en un sombrero: una es roja por ambos lados; otra es negra por ambos lados, y la última es roja por un lado y negra por el otro. Las tres cartas se barajan y se mezclan dentro del sombrero y, posteriormente, se extrae una sin mirarla y se coloca encima de una mesa. Si el lado que se ve es rojo, ¿cuál es la probabilidad de que el otro lado sea negro?

7. En una ciudad, el 52% de los residentes con edad de votar son republicanos, y el otro 48% son demócratas. Entre los residentes, un 64% de los republicanos y un 42% de los demócratas se muestran a favor de que se suspenda una política activa de alquileres promovida por el ayuntamiento. Se selecciona aleatoriamente a un residente con derecho a voto.

   (a) ¿Cuál es la probabilidad de que la persona elegida esté a favor de que se suspenda la política de alquileres?

   (b) Si la persona elegida está en contra de que se suspenda dicha política de alquileres, ¿cuál es la probabilidad de que sea republicana?

8. Un par de genes determinan el color de ojos de una persona. Si ambos genes son de ojos azules, la persona tendrá efectivamente los ojos azules; si ambos genes son de ojos marrones, la persona tendrá efectivamente los ojos marrones; y si un gen es de ojos azules y el otro de ojos marrones, la persona tendrá los ojos marrones. (Por este último hecho se dice que el gen de ojos marrones es *dominante* sobre el gen de ojos azules.) Un recién nacido recibe independientemente un gen de coloración de ojos de cada uno de sus progenitores, y el gen que recibe de cada progenitor será, con las mismas probabilidades, uno de los dos que tiene el progenitor. Suponga que Susana tiene los ojos azules y que sus dos progenitores tienen los ojos marrones.

   (a) ¿Qué par de genes tiene la madre de Susana? ¿Y su padre?

   (b) La hermana de Susana tiene los ojos marrones y está embarazada. Si el marido de su hermana tiene los ojos azules, ¿cuál es la probabilidad de que su hijo tenga los ojos azules? (*Sugerencia*: ¿Cuál es la probabilidad de que la hermana de Susana tenga un gen de ojos azules?)

*4.7 Principios de recuento

Como se ha visto en la sección 4.4, en ocasiones las probabilidades se calculan contando el número de resultados diferentes que caen dentro de un suceso determinado. La clave para que esto se haga de manera efectiva es utilizar la regla conocida como *principio básico de recuento.*

Principio básico de recuento

Supongamos que un experimento consta de dos partes. Si en la parte 1 se pueden obtener n posibles resultados y si, por cada resultado de la parte 1, existen m resultados posibles de la parte 2, el número total de resultados posibles del experimento es nm.

Que este principio básico es cierto se puede ver fácilmente con sólo enumerar todos los resultados posibles del experimento:

$$\begin{array}{llll} (1,1), & (1,2), & \ldots, & (1,m) \\ (2,1), & (2,2), & \ldots, & (2,m) \\ \vdots & & & \\ (n,1), & (n,2), & \ldots, & (n,m) \end{array}$$

donde el resultado del experimento (i, j) significa que en la parte 1 del experimento se ha obtenido el resultado i-ésimo y que en la parte 2 se ha obtenido el resultado j-ésimo. Puesto que la tabla anterior de resultados contiene n filas y cada fila tiene m resultados posibles, existe un total de $m + m \cdots + m = nm$ resultados.

Ejemplo 4.18 Se selecciona aleatoriamente a una mujer y a un hombre de un grupo compuesto por 12 mujeres y 8 hombres. ¿Cuántas elecciones diferentes son posibles?

Solución Si se considera la elección de la mujer como la primera parte del experimento y la elección del hombre como la segunda parte, se desprende del principio básico que existen $12 \cdot 8 = 96$ resultados posibles.

Ejemplo 4.19 Se van a seleccionar a dos personas de un grupo formado por 10 parejas casadas. ¿Cuántas elecciones diferentes son posibles? Si todas las elecciones son igualmente probables, ¿cuál es la probabilidad de que las dos personas seleccionadas estén casadas?

Solución Puesto que la primera persona seleccionada es una de las 20 y que la segunda selección es una de las 19 restantes, del principio básico se deduce que existen $20 \cdot 19 = 380$

resultados posibles. Ahora bien, por cada matrimonio existen dos resultados en los que el matrimonio citado puede ser seleccionado. Estos son que el marido haya sido seleccionado en primer lugar y su esposa en segundo lugar, y al contrario. Así pues, existen $2 \cdot 10 = 20$ resultados diferentes en los que se puede seleccionar a un matrimonio. De aquí resulta que, si se asume que todos los posibles resultados son igualmente probables, la probabilidad de que las personas seleccionadas sean un matrimonio es de $20/380 = 1/19$. ■

Cuando el experimento tenga más de dos partes, el principio básico puede generalizarse como sigue:

Principio básico de recuento generalizado

Supongamos que un experimento consta de r partes. Supongamos que existen n_1 resultados posibles en la parte 1, n_2 resultados posibles en la parte 2, n_3 resultados posibles en la parte 3, y así sucesivamente. En estas condiciones, existen un total de $n_1 \cdot n_2 \cdots n_r$ resultados posibles del experimento.

Como aplicación del principio generalizado, supongamos que se quiere determinar el número de formas diferentes de colocar las tres letras *a, b* y *c* en línea. Por enumeración se puede ver directamente que existen 6 posibilidades:

$$abc,\ acb,\ bac,\ bca,\ cab,\ cba$$

Se puede obtener este mismo resultado si se usa el principio básico de recuento generalizado. Esto es, existen 3 elecciones para la primera letra, existen 2 para la segunda y, por último, existe solamente una elección posible para la tercera. En consecuencia, existen $3 \cdot 2 \cdot 1 = 6$ resultados posibles.

Supongamos ahora que se desea determinar el número de formas en las que se pueden colocar n objetos en fila. Por un razonamiento similar, se ve que existen un total de

$$n \cdot (n-1) \cdot (n-2) \cdots 3 \cdot 2 \cdot 1$$

colocaciones diferentes. Cada una de estas colocaciones determina una *permutación*. Es conveniente introducir la notación $n!$, léase "factorial de n", para representar la expresión anterior. Esto es,

$$n! = n \cdot (n-1) \cdot (n-2) \cdots 3 \cdot 2 \cdot 1$$

Así, por ejemplo,

$$1! = 1$$

$$2! = 2 \cdot 1 = 2$$

$$3! = 3 \cdot 2 \cdot 1 = 6$$

$$4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$$

y así sucesivamente. Además, es conveniente, por definición, asignar a 0! el valor 1.

Ejemplo 4.20 Si en una habitación hay 4 personas, ¿cuál es la probabilidad de que dos de ellas no hayan nacido el mismo día del año?

Solución Puesto que cada persona puede haber nacido en cualquiera de los 365 días del año, se desprende del principio generalizado que existen un total de

$$365 \cdot 365 \cdot 365 \cdot 365 = (365)^4$$

resultados posibles. (Se omite la posibilidad de que alguno haya nacido un 29 de febrero.) Determinemos ahora el número de resultados en los que no existan dos personas que hayan nacido el mismo día del año. Esto ocurrirá si la primera persona ha nacido un día cualquiera de los 365 del año, la segunda ha nacido un día cualquiera de los 364 restantes, la tercera persona ha nacido uno cualquiera de los 363 días restantes, y, finalmente, la última persona ha nacido cualquier día de los restantes 362. Así pues, por el principio básico de recuento generalizado, se ve que existen un total de

$$365 \cdot 364 \cdot 363 \cdot 362$$

resultados posibles; en los que los nacimientos de las 4 personas hayan ocurrido en días distintos. Por ello se deduce que, si todos los resultados posibles son igualmente probables, la probabilidad de que ningún par de personas haya nacido el mismo día es:

$$\frac{365 \cdot 364 \cdot 363 \cdot 362}{365 \cdot 365 \cdot 365 \cdot 365} = 0{,}983644$$

Este mismo razonamiento se puede emplear para calcular la probabilidad de que todos los miembros de un grupo de n personas hayan nacido en días diferentes, para cualquier entero n. Como curiosidad, se puede comprobar que si $n = 23$, esta probabilidad es menor que 1/2. Esto es, si hay 23 personas en la habitación, el hecho de que al menos dos cumpleaños coincidan es más probable que el hecho de que no coincida ningún par de ellos. ■

Ahora suponga que estamos interesados en elegir 3 de las 5 letras *a, b, c, d* y *e*. ¿Cuántas elecciones diferentes son posibles? Para contestar a esta pregunta, se puede razonar que, puesto que existen 5 posibilidades para la primera elección, 4 posibilidades para la segunda y 3 para la tercera, se tiene que existen $5 \cdot 4 \cdot 3$ posibles elecciones, cuando el orden de elección se considera relevante. Sin embargo, en este conjunto de elecciones ordenadas, cada grupo de tres letras aparece 3! veces. Por ejemplo, si consideramos el grupo de las letras *a, b* y *c*, cada una de las permutaciones

abc, acb, bac, bca, cab, cba

de estas tres letras estarán incluidas en el conjunto de elecciones posibles cuando cuente el orden de selección. En consecuencia, resulta que el número de grupos diferentes de tamaño 3 que se pueden formar con las 5 letras, cuando se considera que el orden de selección no tiene importancia, es

$$\frac{5 \cdot 4 \cdot 3}{3 \cdot 2 \cdot 1} = 10$$

Supongamos ahora que se está interesado en determinar el número de grupos diferentes de tamaño r que se pueden extraer de un conjunto de n elementos. Por un razonamiento similar al anterior, se tiene que existen

$$\frac{n \cdot (n-1) \cdots (n-r+1)}{r!}$$

grupos diferentes. Puesto que $n(n-1)\cdots(n-\mathrm{r}+1)$ se puede escribir como $n!/(n-r)!$, el valor anterior se puede expresar como $n!/[(n-r)!r!]$.

Notación y terminología

Definamos $\binom{n}{r}$, para $r \leq n$, mediante

$$\binom{n}{r} = \frac{n!}{(n-r)!r!} = \frac{n(n-1)\cdots(n-r+1)}{r!}$$

La expresión $\binom{n}{r}$ se denomina número de combinaciones de n elementos tomados de r en r; representa el número de grupos distintos de tamaño r que se pueden extraer de un conjunto de n elementos cuando el orden de selección no tiene importancia.

Ejemplo 4.21

(a) ¿Cuántos grupos de tamaño 2 se pueden hacer con los elementos *a, b* y *c*?

(b) ¿Cuántos grupos de tamaño 2 se pueden extraer de un conjunto de 6 personas?

(c) ¿Cuántos grupos de tamaño 3 se pueden extraer de un conjunto de 6 personas?

Solución

(a) Existen $\binom{3}{2} = \frac{3 \cdot 2}{2 \cdot 1} = 3$ grupos diferentes con 2 elementos que se pueden extraer del conjunto *a, b, c*: *a* y *b*, *a* y *c*, *b* y *c*.

(b) y (c) De un conjunto de 6 personas se pueden extraer

$$\binom{6}{2} = \frac{6 \cdot 5}{2 \cdot 1} = 15$$

grupos diferentes de tamaño 2, y

$$\binom{6}{3} = \frac{6 \cdot 5 \cdot 4}{3 \cdot 2 \cdot 1} = 20$$

grupos diferentes de tamaño 3. ■

Ejemplo 4.22 Se extrae una muestra aleatoria de tamaño 3 de un conjunto de 10 elementos. ¿Cuál es la probabilidad de que un elemento prefijado caiga en la muestra?

Solución Existen $\binom{10}{3}$ grupos diferentes que pueden ser elegidos. El número de grupos distintos que contienen el elemento prefijado es igual al número de elecciones de los dos elementos adicionales que se pueden extraer de los 9 elementos restantes, tras haber elegido el elemento prefijado. Así pues, existen $\binom{9}{2}$ grupos distintos que contienen el elemento dado. Así, asumiendo que una muestra aleatoria es aquella en la que cada grupo tiene la misma probabilidad de ser seleccionado, se ve que la probabilidad de que un elemento concreto pertenezca a la muestra es

$$\frac{\binom{9}{2}}{\binom{10}{3}} = \frac{9 \cdot 8}{2 \cdot 1} \div \frac{10 \cdot 9 \cdot 8}{3 \cdot 2 \cdot 1} = \frac{9 \cdot 8 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot 10 \cdot 9 \cdot 8} = 3/10$$

Esto es, hay 3 posibilidades sobre 10 de que un elemento dado pertenezca a la muestra. ■

Ejemplo 4.23 Se ha de seleccionar un comité de 4 personas entre un grupo de 5 hombres y 7 mujeres. Si la selección se hace aleatoriamente, ¿cuál es la probabilidad de que el comité esté compuesto por 2 hombres y 2 mujeres?

Solución Se asumirá que "la selección se hace aleatoriamente", lo que significa que cada una de las $\binom{12}{4}$ combinaciones posibles tiene la misma probabilidad de ser elegida. Puesto que existen $\binom{5}{2}$ elecciones posibles de 2 hombres y $\binom{7}{2}$ elecciones posibles de 2 mujeres, se desprende del principio básico de recuento que existen $\binom{5}{2}\binom{7}{2}$ resultados posibles que contengan 2 hombres y 2 mujeres. En consecuencia, la probabilidad pedida será

$$\frac{\binom{5}{2}\binom{7}{2}}{\binom{12}{4}} = \frac{5 \cdot 4 \cdot 7 \cdot 6 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot 2 \cdot 1 \cdot 12 \cdot 11 \cdot 10 \cdot 9} = \frac{14}{33}$$

Se sigue de la fórmula

$$\binom{n}{r} = \frac{n!}{r!(n - r)!}$$

que

$$\binom{n}{r} = \binom{n}{n-r} \quad \blacksquare$$

Ejemplo 4.24 Compare $\binom{8}{5}$ y $\binom{12}{10}$.

Solución

$$\binom{8}{5} = \binom{8}{3} = \frac{8 \cdot 7 \cdot 6}{3 \cdot 2 \cdot 1} = 56$$

$$\binom{12}{10} = \binom{12}{2} = \frac{12 \cdot 11}{2 \cdot 1} = 66 \quad \blacksquare$$

Se puede comprobar la igualdad $\binom{n}{r} = \binom{n}{n-r}$ mediante un "razonamiento de recuento". Supongamos que pretendemos seleccionar r elementos de un conjunto de n elementos. Puesto que esto se puede hacer seleccionando directamente los r elementos del grupo, o bien seleccionando los $n - r$ elementos que no pertenecerán al grupo, se tiene que el número de elecciones de r elementos es igual al número de elecciones de $n - r$ elementos. Por ejemplo, cualquier elección de 8 cualquiera de los 10 primeros números naturales se corresponde con una elección de los 2 enteros que no están entre los 8 anteriores.

Ejemplo 4.25 Supongamos que $n + m$ dígitos se van a colocar en fila: n son iguales a 1 y m son iguales a 0. ¿Cuántas colocaciones son posibles? Por ejemplo, si $n = 2$ y $m = 1$, existen tres posibles colocaciones:

$$1, 1, 0 \qquad 1, 0, 1 \qquad 0, 1, 1$$

Solución Cada colocación tiene un dígito en la posición 1, otro dígito en la posición 2, otro en la 3, . . . , y finalmente otro en la posición $n + m$. Se puede describir cada colocación si se especifican las n posiciones ocupadas por los dígitos 1. Así, cada elección distinta de n de las $n + m$ posiciones producirá una colocación diferente. Por consiguiente, existen $\binom{n+m}{n}$ colocaciones distintas.

Naturalmente, también se podría describir cada colocación si se especifican las m posiciones ocupadas por los dígitos 0. Esto nos llevaría a la solución $\binom{n+m}{m}$, que es igual a la anterior $\binom{n+m}{n}$. $\blacksquare$

Problemas

1. Si las 3 primeras posiciones están ocupadas por letras y las 4 últimas posiciones lo están por números, ¿cuántas matrículas diferentes de 7 posiciones se pueden obtener?

2. ¿Cuántos posibles órdenes de bateo tiene un equipo de béisbol con 9 jugadores?

3. $9! = 362\,880$. ¿Cuál es el valor de 10!?

4. Existe un tipo de cierre de seguridad por clave que consiste en una rueda para especificar la clave que puede pararse en una de las 36 posiciones existentes, numeradas del 1 al 36. Para abrir el cierre se debe: i) girar la rueda en el sentido de las agujas de un reloj hasta que se alcance cierto número, ii) girar la rueda en el sentido contrario al de las agujas de un reloj hasta alcanzar un segundo número, y iii) volver a girar la rueda en el sentido de las agujas de un reloj hasta alcanzar un tercer número. Si se han olvidado los tres números citados que forman la clave de apertura (que no necesariamente tienen por qué ser distintos), ¿cuántas posiciones distintas se podrían tener que probar para conseguir que el cierre se abra?

5. Los códigos de área de los teléfonos de Estados Unidos y Canadá se componen de una sucesión de tres dígitos: el primero es un entero entre el 2 y el 9; el segundo es el 0 o el 1; y el tercer dígito es un entero entre el 1 y el 9. ¿Cuántos códigos de área son posibles? ¿Cuántos códigos de área que comiencen por 4 pueden existir?

6. Un conocido cuento de niños pone en palabras de un viajero lo siguiente:

   Dirigiéndome a San Ives
   me encontré a un hombre con 7 mujeres.
   Cada mujer tenía 7 sacos.
   Cada saco tenía 7 gatos.
   Cada gato tenía 7 gatitos.
   ¿Con cuántos gatitos se encontró el viajero?

7. (a) Si se deben asignar 4 trabajadores a 4 trabajos, ¿cuántas asignaciones distintas existen?

   (b) ¿Cuántas asignaciones se pueden hacer si los trabajadores 1 y 2 sólo están cualificados para realizar los trabajos 1 y 2, y los trabajadores 3 y 4 lo están para hacer solamente los trabajos 3 y 4?

8. Utilice la fórmula

$$\binom{n}{r} = \frac{n!}{(n-r)!r!}$$

   para encontrar $\binom{n}{0}$, donde n es un entero positivo. Recuerde que 0! es, por definición, igual a 1. Puesto que $\binom{n}{r}$ se supone que es igual al número de grupos de tamaño r que se pueden extraer de un conjunto de n objetos, ¿cree que la respuesta tiene sentido?

9. Calcule los valores de las expresiones siguientes:

$$\binom{8}{4}, \quad \binom{9}{2}, \quad \binom{7}{6}, \quad \binom{10}{3}$$

10. Considere a un grupo de 20 personas. Si cada una saluda dando la mano a todos los demás, ¿cuántos saludos se realizan?

11. Un estudiante debe elegir cuatro asignaturas entre Francés, Español, Historia, Física y Literatura.

    (a) ¿Cuántas elecciones posibles se pueden hacer?

    (b) Si el estudiante hace su elección aleatoriamente, ¿cuál es la probabilidad de que elija simultáneamente Francés y Español?

12. Una compañía de transporte tiene 10 camiones, de los cuales 3 tienen fallos de frenos. Si un inspector elige aleatoriamente 2 de los camiones y chequea sus frenos, ¿cuál es la probabilidad de que ninguno de los camiones elegidos tenga fallos de freno?

13. Una compañía recibe regularmente una gran cantidad de suministros de chips de ordenador. La política de la compañía consiste en elegir aleatoriamente y comprobar 10 de los chips. Si 2 o más de éstos son defectuosos, el suministro se devuelve; en otro caso, el suministro se acepta. Supongamos que un suministro de 100 chips contiene 14 que son defectuosos.

    (a) ¿Cuál es la probabilidad de que la muestra inspeccionada no contenga ningún chip defectuoso?

    (b) ¿Cuál es la probabilidad de que la muestra inspeccionada no contenga 1 chip defectuoso?

    (c) ¿Cuál es la probabilidad de que el suministro sea rechazado?

14. En una lotería estatal, el jugador debe elegir 8 números del 1 al 40. La Comisión de Lotería lleva a cabo, después, una selección de 8 de los citados 40 números. Si en la extracción de la Comisión las $\binom{40}{8}$ combinaciones son igualmente probables, encuentre la probabilidad de que un jugador tenga:

    (a) Los mismos 8 números que los extraídos por la Comisión.

    (b) 7 números iguales a los de ésta.

    (c) Al menos 6 números iguales a los de ésta.

15. La lista aprobada de posibles miembros de un jurado popular contiene 22 hombres y 18 mujeres. Si el jurado ha de estar formado por 12 miembros, encuentre la probabilidad de que, en una selección aleatoria de 12 de las personas de la lista, el jurado resultante esté formado por:

    (a) 22 hombres y 18 mujeres

    (b) 8 mujeres y 4 hombres

    (c) al menos 10 hombres

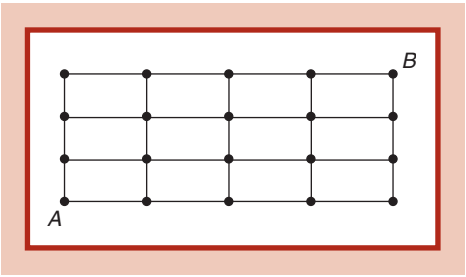
16. Se cuenta que el segundo conde de Yarborough llegó a apostar 1000 contra 1 a que una mano de bridge de 13 cartas contiene al menos una carta que sea 10 o más. (10 o más significa que sea 10, sota, reina, rey o as.) Hoy día, una mano que no tenga cartas superiores a 9 se denomina *Yarborough*. ¿Cuál es la probabilidad de que una mano de bridge seleccionada aleatoriamente sea un Yarborough?

17. Una profesora da a sus alumnos un conjunto de 10 problemas e indica que el examen final (una semana después) constará de una selección aleatoria de 5 de ellos. Si un estudiante sabe resolver 7 problemas en la fecha del examen, encuentre la probabilidad de que conteste correctamente:

    (a) los 5 problemas del examen

    (b) al menos 4 problemas del examen

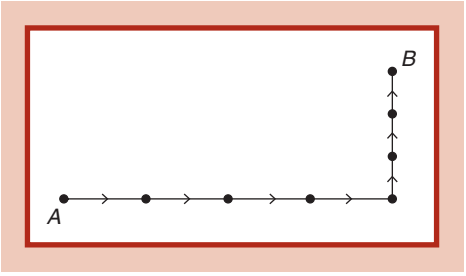
18. Considere la malla de punto mostrada a continuación.



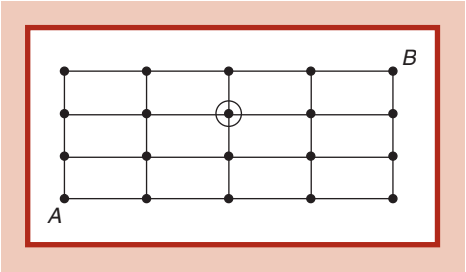
Supongamos que, si se empieza en el punto A, uno se puede mover un paso hacia arriba o un paso hacia la derecha. ¿Cuántos caminos parten de A y llegan hasta B? *Sugerencia:* Para ir de A a B, se deben dar 4 pasos hacia la derecha y 3 hacia arriba. Por consiguiente, cualquiera de los caminos citados viene especificado mediante una colocación particular de 4 letras d y 3 letras a. Por ejemplo, la colocación

$$d, d, d, d, a, a, a$$

representaría el camino siguiente.



19. Supongamos que en el problema 18 se elige aleatoriamente uno de los caminos entre A y B. ¿Cuál es la probabilidad de que el camino elegido pase por el punto marcado con un círculo en la malla siguiente? (*Sugerencia:* ¿Cuántos caminos hay desde A hasta el punto marcado? ¿Cuántos hay desde el punto marcado hasta B?)



Términos clave

Experimento: Cualquier proceso que produzca una observación.

Resultado: La observación producida por un experimento.

Espacio muestral: Conjunto de todos los resultados posibles de un experimento.

Suceso: Cualquier conjunto de resultados de un experimento. Un suceso es un subconjunto del espacio muestral S. Se dice que un suceso ocurre si contiene el resultado del experimento.

Unión de sucesos: La unión de los sucesos A y B, denotada por $A \cup B$, consiste en todos los resultados que están en A o en B o en ambos.

Intersección de sucesos: La intersección de los sucesos A y B, denotada por $A \cap B$, consiste en todos los resultados que están tanto en A como en B.

Complementario de un suceso: El complementario de un suceso A, denotado por A^c, consiste en todos los resultados que no están en A.

Sucesos mutuamente excluyentes o disjuntos: Varios sucesos son mutuamente excluyentes o disjuntos si no pueden ocurrir simultáneamente.

Suceso nulo: Suceso que no contiene ningún resultado. Es el complementario del espacio muestral S.

Diagrama de Venn: Una representación gráfica de los sucesos.

Probabilidad de un suceso: La probabilidad de un suceso A, denotada por $P(A)$, es la probabilidad de que A contenga el resultado del experimento.

Regla de adición de la probabilidad: La fórmula

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Probabilidad condicionada: Probabilidad de un suceso cuando se sabe que ha ocurrido un segundo suceso. Si ha ocurrido A, la probabilidad condicionada de B se denota por $P(B/A)$.

Regla de la multiplicación: La fórmula

$$P(A \cap B) = P(A)P(B|A)$$

Independencia: Dos sucesos son independientes si el conocimiento de que uno de ellos haya o no ocurrido no modifica la probabilidad de que ocurra el otro.

Resumen

Si el resultado de un experimento no es predecible de antemano, S denota el conjunto de todos los posibles resultados del mismo. S se denomina *espacio muestral* del experimento.

A cualquier conjunto de resultados, o equivalentemente a cualquier subconjunto de S, se le llama *suceso*. Si A y B son sucesos, $A \cup B$ (llamado *unión* de A y B) es el suceso com-

puesto por todos los resultados que están en A, en B o en ambos. El suceso $A \cap B$ se llama *intersección* de A y B. Consiste en todos los resultados que están tanto en A como en B.

Para cualquier suceso A, se define el suceso A^c (llamado *complementario* de A) como aquel que contiene todos los resultados de S que no están en A. El suceso S^c, que no contiene ningún resultado, se designa por $\emptyset$. Si $A \cap B = \emptyset$, lo que significa que A y B no tienen elementos comunes, se dice que A y B son *disjuntos* (también llamados *mutuamente excluyentes*).

Se supondrá que, para cada suceso A, existe un número $P(A)$, llamado *probabilidad de A*. Estas probabilidades satisfacen las siguientes tres propiedades:

PROPIEDAD 1: $0 \leq P(A) \leq 1$

PROPIEDAD 2: $P(S) = 1$

PROPIEDAD 3: $P(A \cup B) = P(A) + P(B)$ cuando $A \cap B = \emptyset$

El valor $P(A)$ representa la probabilidad de que el resultado del experimento esté en A. Cuando es así, se dice que A ocurre.

La identidad

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

se denomina *regla de adición de la probabilidad.*

En ocasiones se asume que todos los resultados del experimento son igualmente probables. Bajo esta hipótesis, se puede demostrar que

$$P(A) = \frac{\text{Número de resultados en } A}{\text{Número de resultados en } S}$$

La probabilidad condicionada de B dado que A ocurrió se denota por $P(B/A)$ y viene dada por la siguiente ecuación:

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Si los dos miembros de esta ecuación se multiplican por $P(A)$, se obtiene la igualdad que aparece a continuación, conocida como *regla de la multiplicación*:

$$P(A \cap B) = P(A)P(B|A)$$

Si

$$P(A \cap B) = P(A)P(B)$$

se dice que los sucesos A y B son *independientes*. Si A y B son independientes, la probabilidad de que uno de ellos ocurra no se ve alterada por la información de que el otro haya ocurrido o no.

Problemas de repaso

1. De las 12 botellas de una caja de vino, 3 están mal. Supongamos que se extraen aleatoriamente dos botellas de la caja. Encuentre la probabilidad de que:

   (a) La primera botella elegida esté bien.

   (b) La segunda esté bien.

   (c) Ambas botellas estén bien.

   (d) Las dos botellas estén mal.

   (e) Una esté bien y otra esté mal.

2. Una jugadora de baloncesto encesta cada uno de sus lanzamientos de falta con una probabilidad de 0,8. Supongamos que le hacen una falta y que se la compensa con dos lanzamientos. Asumiendo que los resultados de los distintos lanzamientos de falta son independientes, encuentre la probabilidad de que ella:

   (a) Enceste en los dos lanzamientos.

   (b) Falle en los dos lanzamientos.

   (c) Enceste en el segundo, dado que falló en el primero.

3. Supongamos que una jugadora de baloncesto encesta en su primer lanzamiento de falta con una probabilidad de 0,8. Sin embargo, suponga que su segundo lanzamiento depende de si ha tenido éxito en el primero. Concretamente: si tuvo éxito en el primer lanzamiento, su segundo lanzamiento también lo tiene con una probabilidad de 0,85; mientras que si el primer lanzamiento es fallido, el segundo tiene éxito con una probabilidad de 0,7. Encuentre la probabilidad de que la jugadora:

   (a) Enceste en los dos lanzamientos.

   (b) Falle en los dos lanzamientos.

   (c) Enceste el primero pero falle en el segundo.

4. De los votantes registrados en una comunidad, un 54% son mujeres y un 46% son hombres. El 68% de las mujeres registradas y el 62% de los varones registrados realmente votaron en las últimas elecciones locales. Si se elige aleatoriamente a un votante registrado en esta comunidad, calcule la probabilidad de que dicha persona sea:

   (a) Una mujer que haya votado en las últimas elecciones.

   (b) Un hombre que no haya votado en las últimas elecciones.

   (c) Si esa persona ha votado en las últimas elecciones, ¿cuál es la probabilidad condicionada de que sea un hombre?

5. Una escuela infantil tiene 24 estudiantes: 13 chicas y 11 chicos. Cada día se elige a uno de ellos como “el estudiante del día”. La selección se realiza como sigue. Al comienzo del año académico, los nombres de los niños se escriben en trozos de papel, que se

introducen después en una urna grande. El primer día de escuela, se revuelve la urna y se extrae el nombre del estudiante que va a ser el estudiante del día. Al día siguiente el proceso se repite con los 23 trozos de papel restantes, y así sucesivamente. Cuando cada estudiante ha sido seleccionado una vez (lo cual ocurre en el día 24), se repite el proceso.

(a) ¿Cuál es la probabilidad de que la primera selección sea un chico?

(b) Si la primera selección es un chico, ¿cuál es la probabilidad de que la segunda selección sea una chica?

6. Se seleccionan dos cartas de una baraja de 52. Encuentre la probabilidad de que:

(a) Ambas sean ases.

(b) Ambas sean espadas.

(c) Ambas sean de palos diferentes.

(d) Ambas tengan una puntuación distinta.

7. Si se realizan 6 lanzamientos independientes de una moneda bien construida, ¿cuál es la probabilidad de los resultados siguientes?

(a) C C C C C C

(b) C Z C Z C Z

(c) Z Z C C Z C

8. Calcule la probabilidad de que se conteste perfectamente a varias preguntas de un test con contestaciones de verdadero/falso cuando se eligen aleatoriamente al azar las contestaciones de:

(a) 2 preguntas (b) 3 preguntas

(c) 10 preguntas

9. Una cafetería ofrece un menú con tres platos. Cada cliente elige un plato principal, un acompañamiento y un postre. Las elecciones posibles de cada uno se muestran en la tabla siguiente.

Plato	**Elecciones**
Plato principal	Pollo o filete
Acompañamiento	Arroz o patatas
Postre	Melón o helado o gelatina

Supongamos que el resultado del experimento es el menú completo elegido por un cliente.

(a) Liste todos los resultados contenidos en el espacio muestral.

(b) Si una persona es alérgica al arroz y al melón. Liste todos los resultados del suceso correspondiente a que su elección sea adecuada.

(c) Si una persona elige aleatoriamente un postre, ¿cuál es la probabilidad de que elija el helado?

(d) Si una persona elige aleatoriamente cada uno de los platos, ¿cuál es la probabilidad de que elija pollo, arroz y melón?

10. La siguiente es una clasificación de la población de Estados Unidos por edad y sexo. Los números que aparecen en cada una de las clases están en unidades de millón.

	Sexo	
Edad	**Mujeres**	**Hombres**
Hasta 25 años	48,8	50,4
Más de 25 años	74,5	66,6

Supongamos que se elige a una persona aleatoriamente. Si A es el suceso de que la persona sea varón y B el suceso de que su edad no sobrepase los 25 años, encuentre:

(a) $P(A)$ y $P(A^c)$ (b) $P(B)$ y $P(B^c)$

(c) $P(A \cap B)$ (d) $P(A \cap B^c)$

(e) $P(A|B)$ (f) $P(B|A)$

11. De las tres llaves que tiene una persona, sólo una se ajusta a cierta cerradura. Si prueba las llaves en un orden aleatorio, encuentre la probabilidad de que:

(a) Acierte con la primera llave que prueba.

(b) Acierte con la segunda llave que prueba.

(c) Acierte con la tercera llave que prueba.

(d) Acierte con la segunda llave, dado que no ha acertado con la primera.

12. Una pareja de cartas de una baraja ordinaria constituye un *blackjack* si una es un as y la otra es un 10, una sota, una reina o un rey. Si se extraen dos cartas de la baraja, ¿cuál es la probabilidad de que se obtenga un blackjack? (*Sugerencia:* Se puede intentar calcular la probabilidad de que la primera carta sea un as y la segunda sea un 10, una sota, una reina o un rey, y la probabilidad de que ambas extracciones ocurran al contrario.)

13. Una compañía de transporte dispone de 12 camiones, entre los cuales 4 tienen frenos defectuosos. Si en una inspección, un inspector elige aleatoriamente dos camiones, ¿cuál es la probabilidad de que en ninguno de ellos detecte fallos de frenos?

14. Supongamos que A y B son sucesos independientes, y que

$$P(A) = 0{,}8 \qquad P(B^c) = 0{,}4$$

Encuentre:

(a) $P(A \cap B)$

(b) $P(A \cup B)$

(c) $P(B)$

(d) $P(A^c \cap B)$

15. Se barajan las 52 cartas de una baraja y se van depositando boca arriba una a una.

(a) ¿Cuál es la probabilidad de que la primera carta depositada sea el as de espadas?

(b) Supongamos que el suceso A es que la primera carta depositada no sea el as de espadas y el suceso B es que la segunda carta depositada sea el as de espadas. Por consiguiente, $A \cap B$ es el suceso de que la segunda carta depositada sea el as de espadas. Calcule la probabilidad de este suceso con

$$P(A \cap B) = P(A)P(B|A)$$

(c) Rellene la palabra que falta en el siguiente razonamiento intuitivo para encontrar la solución a la parte (b): Puesto que todas las ordenaciones son igualmente probables, cualquier carta de las 52 tiene una probabilidad _____ de ser la segunda carta depositada boca arriba.

(d) ¿Cuál es la probabilidad de que la 17ª carta puesta boca arriba sea el as de espadas?

16. Distintos discos son sometidos a un plan de inspección que consta de dos etapas. Primero, se chequea cada disco manualmente y, luego, electrónicamente. Si el disco está defectuoso, la inspección manual detecta el error con probabilidad 0,70. Un disco defectuoso que no haya sido detectado manualmente, será detectado electrónicamente con probabilidad 0,80. ¿Qué porcentaje de discos defectuosos no son detectados?

17. Supongamos que las condiciones económicas de un determinado año se pueden clasificar en buenas y malas. Supongamos que, si un año es bueno, el siguiente será también bueno con probabilidad 0,7. De igual forma, si un año es malo, la probabilidad de que el siguiente sea bueno es 0,4. La probabilidad de que este año sea bueno es 0,6. Encuentre las probabilidades de que las sentencias siguientes sean ciertas.

(a) Las condiciones económicas serán buenas tanto este año como el siguiente.

(b) Las condiciones económicas serán buenas este año y serán malas el siguiente.

(c) Las condiciones económicas serán malas los dos años.

(d) Las condiciones económicas serán buenas el año próximo.

(e) Si las condiciones económicas son buenas el año próximo, ¿cuál es la probabilidad condicionada de que las condiciones económicas sean buenas este año?

18. Tanto John como Maureen tienen un gen de ojos azules y un gen de ojos castaños. Un hijo suyo recibirá un gen de color de ojo de Maureen y uno de John. El gen recibido de cada progenitor tiene la misma probabilidad de coincidir con uno de los dos genes de éste. Adicionalmente, el gen que recibe de John es independiente del que recibe de Maureen. Si un hijo recibe un gen de ojos azules tanto de John como de Maureen, el hijo tendrá los ojos azules; en caso contrario, tendrá los ojos castaños. Maureen y John tienen dos hijos.

(a) ¿Cuál es la probabilidad de que su hijo mayor tenga los ojos azules?

(b) ¿Cuál es la probabilidad de que su hijo mayor tenga los ojos azules y el menor los tenga castaños?

(c) ¿Cuál es la probabilidad de que su hijo mayor tenga los ojos castaños y el menor los tenga azules?

(d) ¿Cuál es la probabilidad de que uno de sus hijos tenga los ojos azules y el otro los tenga castaños?

(e) ¿Cuál es la probabilidad de que los dos hijos tengan ojos azules?

(f) ¿Cuál es la probabilidad de que los dos hijos tengan ojos castaños?

19. Se estima que, en la población adulta de Estados Unidos, un 55% está por encima de su peso ideal, un 20% tiene la presión sanguínea alta y un 60% está por encima de su peso ideal o tiene presión sanguínea alta. Si A es el suceso de que una persona adulta elegida aleatoriamente de la población esté por encima de su peso ideal y B es el suceso de que esta persona tenga la presión sanguínea alta, ¿los sucesos A y B son independientes?

20. Se selecciona aleatoriamente una carta de una baraja. A es el suceso de que la carta citada sea un as, y B es el suceso de que sea una espada. Indique si A y B son independientes si se trata de una baraja:

(a) Estándar de 52 cartas.

(b) Estándar, de la que se han sacado las 13 cartas de corazones.

(c) Estándar, de la que se han sacado los corazones con los números comprendidos entre el 2 y el 9.

21. A un total de 500 parejas trabajadoras casadas se les preguntó si cada uno de sus miembros tenía un salario por encima de los 75 000 $. Se obtuvo la siguiente información:

	Marido	
Esposa	**Menos de 75 000 $**	**Más de 75 000 $**
Menos de 75 000 $	212	198
Más de 75 000 $	36	54

Así, por ejemplo, en 36 parejas, la esposa ganaba más de 75 000 $ y el marido menos de esta cantidad. Se selecciona aleatoriamente una de las parejas citadas:

(a) ¿Cuál es la probabilidad de que el marido gane menos de 75 000 $?

(b) Si el marido gana más de 75 000 $, ¿cuál es la probabilidad condicionada de que la esposa gane más de esta cantidad?

(c) Si el marido gana menos de 75 000 $, ¿cuál es la probabilidad condicionada de que la esposa gane más de esta cantidad?

(d) ¿Los salarios de la esposa y el marido son independientes?

22. La probabilidad de que la batería de un coche nuevo funcione durante más de 10 000 millas es 0,8; la probabilidad de que funcione más de 20 000 millas es 0,4, y la proba-

bilidad de que funcione más de 30 000 millas es 0,1. Si una batería de un coche nuevo continúa funcionando tras haber recorrido 10 000 millas, encuentre la probabilidad condicionada de que:

(a) Su vida total exceda a las 20 000 millas.

(b) Su resto de vida exceda a las 20 000 millas.

23. Entre aquellos conductores que se paran en una gasolinera, un 90% adquiere gasolina o gasóleo. Un 86% adquiere gasolina y un 8% adquiere gasóleo.

(a) ¿Qué porcentaje de conductores compran gasolina y gasóleo?

(b) Encuentre la probabilidad condicionada de que un conductor:

(i) Compre gasóleo, dado que ha adquirido gasolina.

(ii) Compre gasolina, dado que ha adquirido gasóleo

(iii) Suponga que un conductor se para en la gasolinera. ¿Los sucesos de que el conductor adquiera gasóleo y de que el conductor adquiera gasolina son independientes?

La tabla siguiente muestra las tasas de participación en distintas actividades artísticas y de ocio de los individuos de una población, por distintas categorías de edad y sexo. Los datos se refieren al año 2000, y las cifras representan la proporción de la población bajo la consideración de que satisface el criterio establecido.

	Asistió al menos una vez a							
Categorías	**Concierto de jazz**	**Concierto de musica clásica**	**Ópera**	**Obra musical**	**Teatro**	**Ballet**	**Visitó al menos una vez un museo o galería de arte**	**Leyó novelas, historias cortas, poesía o teatro**
En promedio	10	13	3	17	12	4	22	56
De 18-24 años	14	11	2	15	11	4	22	57
De 25-34 años	15	12	2	16	12	5	26	59
De 35–44 años	10	16	4	21	14	6	27	62
De 45–54 años	8	15	4	20	13	3	22	57
De 55–64 años	5	11	3	18	10	4	19	50
De 65–74 años	3	13	3	13	10	4	16	50
De 75 años o más	1	10	1	8	7	2	10	48
Hombre	10	11	2	15	11	3	21	48
Mujer	9	14	3	19	12	5	23	63

Fuente: dotación Nacional en Estados Unidos para las Artes.

Los problemas del 24 al 26 se refieren a la tabla citada.

24. Supongamos que se elige aleatoriamente a una persona con una edad comprendida entre los 18 y los 24 años, y a otra con una edad comprendida entre los 35 y los 44 años. Encuentre la probabilidad de que:

    (a) Ambas hayan asistido a un concierto de jazz.

    (b) Sólo una de ellas asistiera a un concierto de jazz.

    (c) Dado que sólo una de ellas ha asistido a un concierto de jazz, ¿cuál es la probabilidad condicionada de que sea la persona más joven de las dos?

25. Supongamos que se selecciona aleatoriamente a un hombre y una mujer. Encuentre la probabilidad de que:

    (a) Sólo uno de ellos asistiera a una representación de ballet.

    (b) Al menos uno de ellos asistiera a una ópera.

    (c) Ambos asistieran a una obra musical.

26. Supongamos que se elige aleatoriamente a un individuo. ¿La tabla proporciona la información suficiente para determinar la probabilidad de que este individuo asistiera tanto a un concierto de jazz como a un concierto de música clásica? Si la contestación es que *no*, ¿bajo qué hipótesis sería posible determinar dicha probabilidad? Bajo esa hipótesis, calcule la probabilidad citada y, después, indique si parece razonable la hipótesis en esta situación.

5

Variables aleatorias discretas

Su sagrada majestad, el azar, lo decide todo.

Voltaire

Se continúa el estudio de la probabilidad con la introducción de las variables aleatorias: magnitudes cuantitativas cuyos valores vienen determinados por el resultado de un experimento. Se definirá el valor esperado de una variable aleatoria, y se analizarán sus propiedades. Se introducirá el concepto de varianza. Se estudiará un tipo muy importante de variables aleatorias, conocidas como *binomiales*.

5.1 Introducción

La Asociación Nacional de Baloncesto (Nacional Basketball Association, NBA) lleva a cabo el siguiente sorteo entre los 11 equipos peores clasificados en el año anterior. Se colocan 66 bolas de ping-pong en una urna. En cada una de estas bolas está escrito el nombre de un equipo; 11 bolas tienen el nombre del equipo peor clasificado, 10 bolas tienen escrito el nombre del segundo equipo con peor clasificación, 9 tienen el nombre del equipo peor clasificado en tercer lugar, y así sucesivamente (hasta llegar a 1 bola con el

nombre del equipo con la 11ª peor clasificación). Después, se elige una bola aleatoriamente, y el equipo cuyo nombre figura en la bola es el primero que tiene derecho de elección sobre la lista de jugadores que entran en la lista de participantes de la liga. Después se sacan todas las bolas que incluyen el nombre de este equipo, y se vuelve a efectuar una extracción. El equipo al que "pertenece" esta bola tiene la segunda opción de elección de jugadores. Finalmente, se elige otra bola, y el equipo marcado recibe la tercera opción. Las restantes opciones, desde la 4ª a la 11ª, se asignan a los 8 equipos "que no ganaron el sorteo", en orden inverso a su clasificación. Por ejemplo, si el equipo peor clasificado no ha recibido ninguna de las tres opciones sorteadas recibirá la cuarta opción de elección de jugadores.

El resultado de este sorteo es el orden con el que los 11 equipos seleccionan a los jugadores. Sin embargo, en lugar de estar interesados en el resultado real, se puede estar interesado en los valores de determinadas magnitudes. Por ejemplo, uno podría estar interesado prioritariamente en saber qué equipo obtiene la primera opción o en saber qué número obtiene el equipo de nuestra ciudad. Estas magnitudes de interés se conocen como *variables aleatorias*, y un tipo especial de éstas, las llamadas *discretas*, se van a estudiar en este capítulo.

Las variables aleatorias se introducirán en la sección 5.2. En la sección 5.3 se introducirá el concepto de valor esperado (o esperanza) de una variable aleatoria. Se verá que éste representa, de forma precisa, el valor medio de la variable aleatoria. Las propiedades de la esperanza se presentan, igualmente, en la sección 5.3.

La sección 5.4 se dedica a la varianza de una variable aleatoria, que es una medida que indica en qué medida la variable aleatoria tiende a diferir de su valor esperado. También, en esta sección, se introducirá el concepto de variables aleatorias independientes.

En la sección 5.5, se trata un tipo muy relevante de variables aleatorias, llamadas *binomiales*. Se verá cómo surgen tales variables aleatorias y se estudiarán sus propiedades.

En las secciones 5.6 y 5.7 se introducirán las variables aleatorias hipergeométricas y de Poisson. Se explicará cómo surgen estas dos variables aleatorias discretas y se estudiarán sus propiedades.

La primera bola extraída de en la lotería antes aludida de la NBA correspondió al equipo Orlando Magic, pese a que éste había terminado la temporada en la 11ª peor clasificada y, por consiguiente, sólo le correspondía una única bola de las 66 de la urna.

5.2 Variables aleatorias

A menudo, cuando se lleva a cabo un experimento aleatorio, no se está interesado en todos los detalles del resultado, sino que, por el contrario, el interés se centra sobre el valor de ciertas magnitudes numéricas determinadas por el mismo. Por ejemplo, cuando se lanzan varios dados, uno puede estar interesado en conocer cuál es la suma obtenida, y no en los resultados concretos obtenidos con cada dado. Igualmente, puede que un inversor no esté interesado en conocer todas las variaciones que se han producido a lo largo del día en el precio de una acción, sino que, por el contrario, sólo le interesa saber el precio al final del día. Estas magnitudes de interés que vienen determinadas por el resultado del experimento se conocen como *variables aleatorias.*

Puesto que el valor de la variable aleatoria depende del resultado el experimento, se pueden asignar probabilidades a sus posibles valores.

Ejemplo 5.1 Si en la lotería de la NBA, citada en la sección 5.1, sólo nos interesan los equipos que obtienen las tres primeras posiciones, los resultados serán ternas de números. Por ejemplo, el resultado (3, 1, 4) podría significar que el equipo que ocupó la tercera peor posición en la temporada recibió la primera opción, el equipo que ocupó la peor posición en la temporada recibió la segunda opción y el equipo que obtuvo la cuarta peor posición en la temporada recibió la tercera opción. Si X denota el equipo que recibió la primera opción, X será igual a 3 si el resultado del experimento ha sido (3, 1, 4).

Claramente, X puede tomar cualquier valor entero entre 1 y 11, ambos inclusive. Será igual a 1 si la primera bola extraída es una de las 11 existentes que corresponde con el equipo que obtuvo la peor posición en la temporada, será igual a 2 si la primera bola extraída es una de las 10 que pertenecen al equipo que en la temporada acabó en la segunda peor posición, y así sucesivamente. Puesto que las 66 bolas tienen la misma probabilidad de ser elegidas en la primera extracción, se sigue que

$$P\{X = 1\} = \frac{11}{66} \qquad P\{X = 7\} = \frac{5}{66}$$

$$P\{X = 2\} = \frac{10}{66} \qquad P\{X = 8\} = \frac{4}{66}$$

$$P\{X = 3\} = \frac{9}{66} \qquad P\{X = 9\} = \frac{3}{66}$$

$$P\{X = 4\} = \frac{8}{66} \qquad P\{X = 10\} = \frac{2}{66}$$

$$P\{X = 5\} = \frac{7}{66} \qquad P\{X = 11\} = \frac{1}{66}$$

$$P\{X = 6\} = \frac{6}{66}$$ ■

Ejemplo 5.2 Supongamos que se desea conocer el sexo de cada uno de los tres hijos de una determinada familia. El espacio muestral de este experimento se compone de los 8 resultados siguientes ($b \equiv$ varón, $g \equiv$ hembra):

$$\{(b, b, b), (b, b, g), (b, g, b), (b, g, g), (g, b, b), (g, b, g), (g, g, b), (g, g, g)\}$$

El resultado (g, b, b) significa, por ejemplo, que: el descendiente más joven es mujer, el siguiente más joven es varón y el mayor es varón. Supongamos que los 8 resultados posibles son igualmente probables y, por tanto, todos ellos tienen probabilidad 1/8.

Si X denota el número de hijas de la familia, el valor de X viene determinado por el resultado del experimento. Esto es, X es una variable aleatoria que puede tomar los valores 0, 1, 2 o 3. Se determinarán a continuación las probabilidades de que X sea igual a cada uno de estos cuatro valores.

Puesto que X es igual a 0, si el resultado es (b, b, b), se ve que

$$P\{X = 0\} = P\{(b, b, b)\} = \frac{1}{8}$$

Puesto que X es igual a 1 si el resultado es (b, b, g) o (b, g, b) o (g, b, b), se tiene que

$$P\{X = 1\} = P(\{(b, b, g), (b, g, b), (g, b, b)\}) = \frac{3}{8}$$

De igual forma

$$P\{X = 2\} = P(\{(b, g, g), (g, b, g), (g, g, b)\}) = \frac{3}{8}$$

$$P\{X = 3\} = P(\{(g, g, g)\}) = \frac{1}{8} \quad \blacksquare$$

Se dice que una variable aleatoria es *discreta* si sus posibles valores forman una sucesión de puntos separados de la recta real. Así pues, por ejemplo, cualquier variable aleatoria que pueda tomar un número finito de valores distintos es discreta.

En este capítulo, se estudiarán las variables aleatorias discretas. Sea X una de ellas y supongamos que puede tomar los n valores posibles: $x_1, x_2, \ldots, x_n$. Como en los ejemplos 5.1 y 5.2, se utilizará la notación $P\{X = x_i\}$ para representar la probabilidad de que X sea igual a x_i. El conjunto de estas probabilidades se denomina *distribución de probabilidad* de X. Puesto que X sólo puede tomar uno de estos n valores, se sabe que

$$\sum_{i=1}^{n} P\{X = x_i\} = 1$$

Ejemplo 5.3 Supongamos que X es una variable aleatoria que puede tomar uno de los valores 1, 2 o 3. Si

$$P\{X = 1\} = 0{,}4 \qquad \text{y} \qquad P\{X = 2\} = 0{,}1$$

¿cuál es el valor de $P\{X = 3\}$?

Solución Puesto que las probabilidades suman 1, se tiene que

$$1 = P\{X = 1\} + P\{X = 2\} + P\{X = 3\}$$

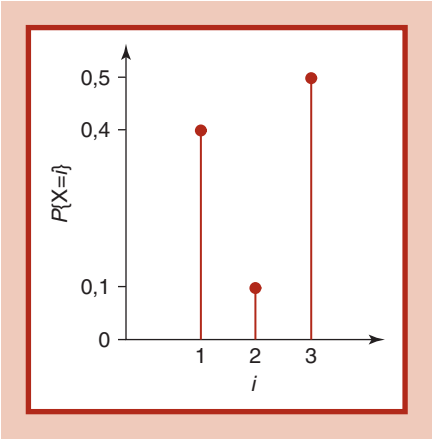


Figura 5.1 Gráfico de $P\{X = i\}$.

o

$$1 = 0{,}4 + 0{,}1 + P\{X = 3\}$$

Por consiguiente,

$$P\{X = 3\} = 1 - 0{,}5 = 0{,}5$$

En la figura 5.1 se muestra el gráfico de $P\{X = i\}$. ■

Ejemplo 5.4 Una vendedora ha concertado dos citas para vender enciclopedias. Cree que en la primera cita puede realizar una venta con una probabilidad de 0,3; que, en la segunda, lo puede hacer con una probabilidad de 0,6, y que los resultados de las dos citas son independientes. ¿Cuál es la distribución de probabilidad de X, el número de ventas realizadas?

Solución La variable aleatoria X puede tomar cualquiera de los valores 0, 1 o 2. Será igual a cero si no se vende en ninguna de las citas; por tanto:

$$P\{X = 0\} = P\{\text{no vende en la primera cita, no vende en la segunda}\}$$

$$= P\{\text{no vende en la primera cita}\}\, P\{\text{no vende en la segunda}\} \text{ por la independencia}$$

$$= (1 - 0{,}3)\,(1 - 0{,}6) = 0{,}28$$

La variable aleatoria X será igual a 1 si consigue vender en la primera cita y no en la segunda, o si no consigue vender en la primera cita y sí lo consigue en la segunda. Puesto que estos dos sucesos son disjuntos, se tiene que:

$$\begin{aligned} P\{X = 1\} &= P\{\text{vende en la primera cita, no vende en la segunda}\} \\ &\quad + P\{\text{no vende en la primera cita, vende en la segunda}\} \\ &= P\{\text{vende en la primera cita}\}P\{\text{no vende en la segunda}\} \\ &\quad + P\{\text{no vende en la primera cita}\}P\{\text{vende en la segunda}\} \\ &= 0{,}3(1 - 0{,}6) + (1 - 0{,}3)0{,}6 = 0{,}54 \end{aligned}$$

Finalmente, la variable aleatoria X será igual a 2 si se realizan ventas en las dos citas; así pues:

$$\begin{aligned} P\{X = 2\} &= P\{\text{vende en la primera cita, vende en la segunda}\} \\ &= P\{\text{vende en la primera cita}\}P\{\text{vende en la segunda cita}\} \\ &= (0{,}3)(0{,}6) = 0{,}18 \end{aligned}$$

Para comprobar los resultados obtenidos, observe que

$$P\{X = 0\} + P\{X = 1\} + P\{X = 2\} = 0{,}28 + 0{,}54 + 0{,}18 = 1 \quad \blacksquare$$

Problemas

1. En el ejemplo 5.2 supongamos que la variable aleatoria Y toma el valor 1 si la familia tiene al menos un hijo de cada sexo, y toma el valor 0 en otro caso. Encuentre $P\{Y = 0\}$ y $P\{Y = 1\}$.
2. En el ejemplo 5.2 supongamos que la variable aleatoria W es igual al número de chicas nacidas antes del primer chico. [Si el resultado es (g, g, g), W toma el valor 3.] Indique los posibles valores de W junto con sus probabilidades. Esto es, obtenga la distribución de probabilidad de W.
3. La tabla siguiente muestra el número total de tornados (columnas de aire, violentas y giratorias, con velocidades de viento superiores a 100 millas por hora) en los Estados Unidos entre 1980 y 1991.

Año	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991
Tornados	866	783	1046	931	907	684	764	656	702	856	1133	1132

Fuente: Administración Nacional Oceánica y Atmosférica de Estados Unidos.

Supongamos que se selecciona aleatoriamente uno de estos años y que X denota el número de tornados de dicho año. Encuentre:

(a) $P\{X > 900\}$

(b) $P\{X \leq 800\}$

(c) $P\{X = 852\}$

(d) $P\{700 < X < 850\}$

4. Supongamos que se lanzan un par de dados. Sea X la suma de los resultados obtenidos. ¿Cuáles son los posibles valores de X? Asumiendo que los 36 resultados posibles del experimento son igualmente probables, ¿cuál es la distribución de probabilidad de X?

5. En el problema 4, denote como Y al menor de los dos resultados obtenidos. (Si las dos caras tienen el mismo valor, asuma que éste es el valor de Y.) Determine la distribución de probabilidad de Y.

6. Dos personas deciden verse en un parque. Se asume que cada persona tiene la misma probabilidad de llegar a las 2:00, a las 2:30 o a las 3:00 del medio día, y que las horas de llegada de ambas personas son independientes. Sea X el tiempo que deberá esperar la primera persona que llegue y asumamos que X toma el valor 0 cuando ambas llegan a la misma hora.

   (a) ¿Cuáles son los posibles valores de X?

   (b) ¿Con qué probabilidades toma X cada uno de los valores anteriores?

7. Dos equipos de voleibol participan en un partido a tres juegos que vence el que primero gane dos de los tres. Supongamos que el equipo de casa gana cada juego, independientemente, con probabilidad 0,7. Denote como X el número de juegos realizados.

   (a) ¿Cuáles son los posibles valores de X?

   (b) ¿Cuál es la distribución de probabilidad de X?

8. Suponga que se extraen aleatoriamente 2 baterías de una caja que contiene 10: 7 que están en buen estado y 3 que son defectuosas. X denota el número de baterías defectuosas elegidas. Determine los posibles valores de X junto con sus respectivas probabilidades.

9. Un suministro consta de 120 piezas, de las cuales 10 son defectuosas. Se eligen aleatoriamente dos de las piezas del suministro y se inspeccionan. X denota el número de piezas que resultan defectuosas. Encuentre la distribución de probabilidad de X.

10. Una mujer solicita dos trabajos sucesivamente. Tiene probabilidad 0,5 de conseguir el primer trabajo solicitado. Si obtiene el primer trabajo, tiene probabilidad 0,2 de conseguir el segundo; mientras que si no obtiene el primer trabajo, la probabilidad de que consiga el segundo es 0,4. (En este último caso, la probabilidad es mayor.) Denote como X el número de trabajos que obtiene. Encuentre la distribución de probabilidad de X.

11. Siempre que un jugador de baloncesto tira dos tiros de falta, acierta en el primero con una probabilidad de 0,75. Si acierta el primero, acierta también el segundo con probabilidad 0,80; mientras que si falla el primero, la probabilidad de que acierte el segundo

es 0,70. Denote como X el número de aciertos por falta. Encuentre la distribución de probabilidad de X.

En los problemas 12, 13 y 14, indique si el conjunto de valores $p(i)$, $i = 1, 2, 3, 4, 5$, puede representar las probabilidades $P\{X = i\}$ de una variable aleatoria cuyos valores posibles son 1, 2, 3 ,4 y 5. Si la respuesta es no, explique por qué.

12.

i	$p(i)$
1	0,4
2	0,1
3	0,2
4	0,1
5	0,3

13.

i	$p(i)$
1	0,2
2	0,3
3	0,4
4	−0,1
5	0,2

14.

i	$p(i)$
1	0,3
2	0,1
3	0,2
4	0,4
5	0,0

15. En un estudio que afecta a 223 hogares de una pequeña ciudad rural de Iowa, un sociólogo ha recogido datos del número de hijos de cada hogar. Los datos muestran que existen 348 hijos en la ciudad, clasificados como sigue: 38 hogares tienen 0 hijos, 82 tienen 1 hijo, 57 tienen 2 hijos, 34 tienen 3 hijos, 10 tienen 4 hijos y 2 tienen 5 hijos. Suponga que se elige aleatoriamente un hogar para llevar a cabo una entrevista detallada. Denótese como X el número de hijos del hogar seleccionado. Obtenga la distribución de probabilidad de X.

16. Suponga que, en el problema 15, se selecciona aleatoriamente a uno de los 348 hijos. Denote como Y el número total de hijos de la familia del hijo seleccionado. Encuentre la distribución de probabilidad de Y.

17. Supongamos que X toma uno de los valores 1, 2, 3, 4 o 5. Si $P\{X < 3\} = 0{,}4$ y $P\{X > 3\} = 0{,}5$ encuentre

    (a) $P\{X = 3\}$

    (b) $P\{X < 4\}$

18. Un agente de seguros tiene dos clientes, cada uno de los cuales ha suscrito una póliza de seguro de vida por la cual se cobrarían 100 000 $ en caso de muerte. Las probabilidades de muerte durante este año para cada uno de los suscriptores son 0,05 y 0,10, respectivamente. Denote como X la cantidad total de dinero que se pagará este año a los beneficiarios de los clientes. Si se asume que las muertes de los clientes 1 y 2 son sucesos independientes, determine la distribución de probabilidad de X.

19. Una pastelería dispone de tres tartas especiales al comienzo del día. La demanda diaria de este tipo de tartas es:

0	con probabilidad 0,15
1	con probabilidad 0,20
2	con probabilidad 0,35
3	con probabilidad 0,15
4	con probabilidad 0,10
5 o más	con probabilidad 0,05

Si X denota el número de tartas que no se han vendido al final del día, determine la distribución de probabilidad de X.

5.3 Valor esperado

Un concepto clave dentro de la Probabilidad es el valor esperado de una variable aleatoria. Si X es una variable aleatoria discreta que puede tomar uno de los valores $x_1, x_2, \ldots, x_n$, el *valor esperado* de X, denotado por $E[X]$, se define por

$$E[X] = \sum_{i=1}^{n} x_i P\{X = x_i\}$$

El valor esperado de X es una media ponderada de los posibles valores de X, en la que el peso de un determinado valor coincide con la probabilidad de que X tome el valor citado. Por ejemplo, supongamos que X toma los valores 0 y 1 con probabilidades iguales, es decir,

$$P\{X = 0\} = P\{X = 1\} = \frac{1}{2}$$

en este caso,

$$E[X] = 0\left(\frac{1}{2}\right) + 1\left(\frac{1}{2}\right) = \frac{1}{2}$$

es igual a la media ordinaria de los dos posibles valores, 0 y 1, que puede tomar X. Por el contrario, si

$$P\{X = 0\} = \frac{2}{3} \quad \text{y} \quad P\{X = 1\} = \frac{1}{3}$$

se tiene que

$$E[X] = 0\left(\frac{2}{3}\right) + 1\left(\frac{1}{3}\right) = \frac{1}{3}$$

es una media ponderada de los dos posibles valores, 0 y 1, donde al valor 0 se le asigna el doble de peso que al valor 1, puesto que la probabilidad de que X sea igual a 0 es el doble de la probabilidad de que X sea igual a 1.

Definición y terminología

El *valor esperado* de una variable aleatoria discreta X, cuyos valores posibles son $x_1, x_2, \ldots, x_n$, se denota como $E[X]$, y viene definido por

$$E[X] = \sum_{i=1}^{n} x_i P\{X = x_i\}$$

Otros nombres utilizados para identificar $E[X]$ son *esperanza* de X y *media* de X.

Otra motivación para la definición del valor esperado se basa en la interpretación frecuentista de las probabilidades. Esta interpretación asume que, si se lleva a cabo un gran número (en teoría, un número infinito) de repeticiones del experimento, la proporción de veces que ocurre el suceso A es igual a $P(A)$. Considere ahora que la variable aleatoria X toma los valores $x_1, x_2, \ldots, x_n$, con probabilidades $p(x_1), p(x_2), \ldots, p(x_n)$, y asuma que X representa las ganancias derivadas de un determinado juego de azar. Razonaremos que, si se realizan un gran número de juegos, la ganancia media por cada juego será $E[X]$. Para ver esto, supongamos que se llevan a cabo N juegos, siendo N muy grande. Puesto que, por la interpretación frecuentista de la probabilidad, la proporción de juegos en los que se gana x_i será aproximadamente $p(x_i)$, se deduce que se ganará x_i en aproximadamente $Np(x_i)$ de los N juegos. Dado que esto es cierto para cada valor x_i, se tiene que las ganancias totales en los N juegos será aproximadamente igual a

$$\sum_{i=1}^{n} x_i \text{ (número de juegos en los que se gana } x_i) = \sum_{i=1}^{n} x_i Np(x_i)$$

Por consiguiente, la ganancia media por juego será

$$\frac{\sum_{i=1}^{n} x_i N p(x_i)}{N} = \sum_{i=1}^{n} x_i p(x_i) = E[X]$$

En otras palabras, si X es una variable aleatoria asociada a un determinado experimento, el valor medio de X, sobre un gran número de repeticiones del experimento, es aproximadamente igual a $E[X]$.

Ejemplo 5.5 Supongamos que se lanza un dado cuyas seis caras tienen la misma probabilidad. Encuentre $E[X]$, donde X denota la cara que resulta en el lanzamiento.

Solución Puesto que

$$P\{X = i\} = \frac{1}{6} \qquad \text{para } i = 1, 2, 3, 4, 5, 6$$

se ve que

$$E[X] = 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) + 6\left(\frac{1}{6}\right)$$

$$= \frac{21}{6} = 3{,}5$$

Observe que el valor esperado de X no coincide con ninguno de los posibles valores de X. Aunque $E[X]$ se denomina valor esperado de X, no se debe interpretar como el valor de X que se espera obtener, sino que, por el contrario, se debe interpretar como el valor medio de X en un gran número de repeticiones del experimento. Es decir, si se lanza el dado sucesivamente, la media de todos los resultados, obtenida sobre un gran número de lanzamientos, será aproximadamente igual a 3,5. ■

Ejemplo 5.6 Considere una variable aleatoria X que toma el valor 1 o el valor 0 con probabilidades p y $1 - p$, respectivamente. Esto es,

$$P\{X = 1\} = p \qquad \text{y} \qquad P\{X = 0\} = 1 - p$$

Encuentre $E[X]$.

Solución El valor esperado de esta variable aleatoria es

$$E[X] = 1(p) + 0\,(1 - p) = p \quad ■$$

Ejemplo 5.7 Una compañía de seguros establece una cuota anual sobre sus seguros de vida de forma que su beneficio esperado coincida con el 1% de la cantidad que debe desembolsar en caso de muerte. Encuentre la cuota anual de una póliza de seguro de vida de 200 000 $ para un individuo cuya probabilidad de muerte durante el año sea 0,02.

Solución En unidades de 1000 $, la compañía de seguros fijará su cuota de forma que el beneficio esperado sea un 1% de 200; es decir, 2. Si A denota la cuota anual, el beneficio de la compañía de seguros será

$$A \qquad \text{si el asegurado vive}$$

o

$$A - 200 \qquad \text{si el asegurado muere}$$

Por consiguiente, el beneficio esperado viene dado por:

$$\begin{aligned} E[\text{beneficio}] &= AP\{\text{de que el asegurado viva}\} + (A - 200)\, P\{\text{de que el asegurado muera}\} \\ &= A\,(1 - 0{,}02) + (A - 200)(0{,}02) \\ &= A - 200(0{,}02) \\ &= A - 4 \end{aligned}$$

Así pues, la compañía obtendrá un beneficio esperado de 2000 $ si establece una cuota anual de 6000 $ al asegurado en cuestión. ■

Como se ha visto en el ejemplo 5.7, $E[X]$ se mide siempre en las mismas unidades (dólares en el anterior ejemplo) en las que se mide la variable aleatoria X.

El concepto de valor esperado es análogo al concepto físico del centro de gravedad de una distribución de masas. Considere una variable aleatoria cuyas probabilidades son $p(x_i)$, $i \geq 1$. Si se imagina una barra en la que se cuelgan pesos $p(x_i)$ en los puntos x_i, $i \geq 1$ (figura 5.2), el punto en el que la barra se encontraría en equilibrio se conoce como *centro de gravedad*. Se puede demostrar que, por las leyes de la mecánica, dicho punto es

$$\sum_i x_i p(x_i) = E[X]$$

5.3.1 Propiedades del valor esperado

Sea X una variable aleatoria X con valor esperado $E[X]$. Si c es una constante, las magnitudes cX y $X + c$ también son variables aleatorias y, por tanto, se podrán calcular sus valores esperados. Se pueden demostrar los resultados siguientes:

$$E[cX] = cE[X]$$

$$E[X + c] = E[X] + c$$

Esto es, el valor esperado de una variable aleatoria multiplicada por una constante es igual al valor esperado de la variable aleatoria multiplicada por la constante; y el valor esperado de una variable aleatoria más una constante es igual al valor esperado de la variable aleatoria más la constante.

Ejemplo 5.8 Una pareja casada trabaja para un empresario. La paga extra de Navidad de la mujer es una variable aleatoria cuyo valor esperado es 1500 dólares.

(a) Si la paga extra del marido se fija igual al 80% de la de la mujer, encuentre el valor esperado de la paga extra del marido.

(b) Si la paga extra del marido se establece igual a 1000 $ más que la de su mujer, encuentre su valor esperado.

Solución Denote como X la paga extra (en dólares) de la mujer.

(a) Puesto que la paga extra del marido es $0{,}8X$, se tiene que

$$E[\text{paga extra del marido}] = E[0{,}8X] = 0{,}8E[X] = 1200\ \$$$

(b) En este caso, la paga extra del marido es $X + 1000$; por consiguiente,

$$E[\text{paga extra del marido}] = E[X + 1000] = E[X] + 1000 = 2500\ \$$$ ■

Una propiedad de gran utilidad es que el valor esperado de la suma de variables aleatorias es igual a la suma de los valores esperados individuales.

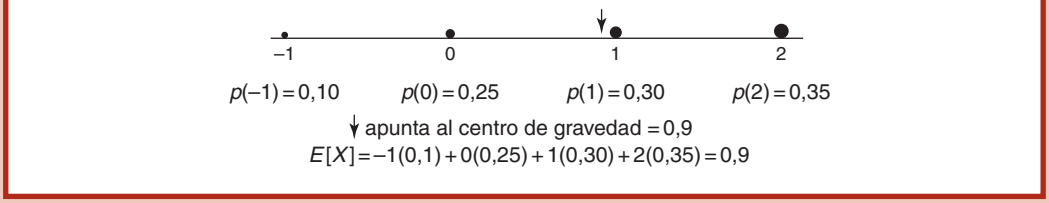


Figura 5.2 Centro de gravedad = $E[X]$.

Para las variables aleatorias X e Y,

$$E[X + Y] = E[X] + E[Y]$$

Ejemplo 5.9 Los datos siguientes representan las rentas anuales de 7 hombres y 7 mujeres que residen en cierta comunidad.

Renta anual (en miles de dólares)

Hombres	Mujeres
33,5	24,2
25,0	19,5
28,6	27,4
41,0	28,6
30,5	32,2
29,6	22,4
32,8	21,6

Supongamos que se elige aleatoriamente a un hombre y una mujer. Encuentre el valor esperado de la suma de sus rentas.

Solución Denote como X la renta del hombre y como Y la renta de la mujer. Puesto que es igualmente probable que X sea cualquiera de los 7 valores que figuran en la columna de los hombres, se ve que

$$E[X] = \frac{1}{7}(33{,}5 + 25 + 28{,}6 + 41 + 30{,}5 + 29{,}6 + 32{,}8)$$

$$= \frac{221}{7} \approx 31{,}571$$

De igual forma,

$$E[Y] = \frac{1}{7}(24{,}2 + 19{,}5 + 27{,}4 + 28{,}6 + 32{,}2 + 22{,}4 + 21{,}6)$$

$$= \frac{175{,}9}{7} \approx 25{,}129$$

Por consiguiente, el valor esperado de la suma de sus rentas es

$$E[X + Y] = E[X] + E[Y]$$

$$\approx 56{,}700$$

Esto es, el valor esperado de la suma de sus rentas es de aproximadamente 56 700 dólares. ■

Ejemplo 5.10 La tabla siguiente muestra el número de guardias de prisiones a tiempo completo en las cárceles de ocho ciudades en 1990.

Ciudad	Guardias de prisiones
Minneapolis, MN	105
Newark, NJ	155
Omaha, NE	149
Portland, OR	195
San Antonio, TX	290
San Jose, CA	357
Tucson, AZ	246
Tulsa, OK	178

Fuente: Departamento de Justicia, *Informe de criminalidad en Estados Unidos, 1990.*

Supongamos que, sucesivamente, se eligen aleatoriamente dos ciudades y que se entrevista a todos sus guardias de prisiones. Encuéntrese el número medio de personas entrevistadas.

Solución X e Y representan los números de guardias de prisiones en la primera y la segunda ciudad, respectivamente. Puesto que la selección de las ciudades es aleatoria, las 8 ciudades tienen la misma probabilidad de resultar elegidas, tanto en la primera opción como en la segunda. Por consiguiente, tanto X como Y pueden tomar cualquier valor de la tabla con unas probabilidades iguales; por consiguiente

$$E[X] = E[Y] = \frac{1}{8}(105 + 155 + 149 + 195 + 290 + 357 + 246 + 178)$$

$$= \frac{1675}{8}$$

de donde,

$$E[X + Y] = E[X] + E[Y] = \frac{1675}{4} = 418{,}75$$

Esto es, el número esperado de entrevistas es de 418,75. ■

Si se utiliza la interpretación frecuentista que hace coincidir el valor esperado con el valor medio de la variable aleatoria, calculado éste sobre un gran número de repeticiones del experimento, es fácil ver intuitivamente por qué el valor esperado de una suma es igual a la suma de los valores esperados. Por ejemplo, supongamos que siempre se hacen las mismas dos apuestas en cada giro de la ruleta, una apuesta referida al color resultante y la otra referida al número resultante. X e Y son las cantidades (en dólares) perdidas en las apuestas sobre el color y sobre el número, respectivamente, en cada giro. Por consiguiente, $X + Y$ representa la pérdida total en cada apuesta doble. Ahora bien, a largo plazo, si se pierde un

promedio de 1 unidad por la apuesta sobre el color (esto es, $E[X] = 1$), y un promedio de 2 unidades en cada apuesta sobre el número (esto es, $E[Y] = 2$), la pérdida total media en cada giro de la ruleta (igual a $E[X + Y]$) será claramente igual a $1 + 2 = 3$.

El resultado de que el valor esperado de una suma de variables aleatorias es igual a la suma de los valores esperados no se verifica únicamente para dos variables aleatorias, sino para cualquier número de ellas.

Resultado de utilidad

Para cualquier entero positivo k y para cualesquiera variables aleatorias $X_1, \ldots, X_k$,

$$E\left[\sum_{i=1}^{k} X_i\right] = \sum_{i=1}^{k} E[X_i]$$

Ejemplo 5.11 Un constructor ha presentado ofertas de construcción para tres obras. Si consigue las tres obtendrá unos beneficios de 20, 25 y 40 (en unidades de 1000 $). Por otro lado, cada obra que no consiga le ocasionará una pérdida (debida al tiempo y al dinero empleado en hacer la oferta) de 2 unidades. Si las probabilidades de que el constructor consiga las obras son 0,3, 0,6 y 0,2, respectivamente, ¿cuál es el beneficio total esperado?

Solución X_i representa el beneficio obtenido por la obra i, $i = 1, 2, 3$. Si se interpretan las pérdidas como beneficios negativos, se tiene

$$P\{X_1 = 20\} = 0{,}3 \quad P\{X_1 = -2\} = 1 - 0{,}3 = 0{,}7$$

Por consiguiente,

$$E[X_1] = 20(0{,}3) - 2(0{,}7) = 4{,}6$$

De la misma forma,

$$E[X_2] = 25(0{,}6) - 2(0{,}4) = 14{,}2$$

y

$$E[X_3] = 40(0{,}2) - 2(0{,}8) = 6{,}4$$

El beneficio total es $X_1 + X_2 + X_3$, de donde

$$\begin{aligned} E[\text{beneficio total}] &= E[X_1 + X_2 + X_3] \\ &= E[X_1] + E[X_2] + E[X_3] \\ &= 4{,}6 + 14{,}2 + 6{,}4 \\ &= 25{,}2 \end{aligned}$$

Así pues, el beneficio total esperado es de 25 200 $. ■

Problemas

En los siguientes problemas, $p(i)$ representa $P\{X = i\}$.

1. Encuentre el valor esperado de X cuando
   (a) $p(1) = 1/3, p(2) = 1/3, p(3) = 1/3$
   (b) $p(1) = 1/2, p(2) = 1/3, p(3) = 1/6$
   (c) $p(1) = 1/6, p(2) = 1/3, p(3) = 1/2$
2. Encuentre $E[X]$ cuando
   (a) $p(1) = 0{,}1, p(2) = 0{,}3, p(3) = 0{,}3, p(4) = 0{,}2, p(5) = 0{,}1$
   (b) $p(1) = 0{,}3, p(2) = 0{,}1, p(3) = 0{,}2, p(4) = 0{,}1, p(5) = 0{,}3$
   (c) $p(1) = 0{,}2, p(2) = 0, \quad p(3) = 0{,}6, p(4) = 0, \quad p(5) = 0{,}2$
   (d) $p(3) = 1$
3. Un distribuidor obtiene un beneficio de 50 dólares por cada artículo que se recibe en perfecto estado y sufre una pérdida de 6 dólares por cada artículo recibido cuyo estado no sea perfecto. Si la probabilidad de que un artículo se reciba en perfecto estado es 0,4, ¿cuál es el beneficio esperado del distribuidor por cada artículo recibido?
4. En un determinado juicio, una abogada debe decidir si cobra una tarifa directa de 1200 $ o si la fija según el resultado, en cuyo caso recibirá 5000 $ solamente si gana el juicio. Determine si es la tarifa directa o la tarifa condicionada la que le proporciona un mayor rendimiento esperado, asumiendo que la probabilidad de ganar el caso es:
   (a) 1/2
   (b) 1/3
   (c) 1/4
   (d) 1/5
5. Suponga que X puede tomar cualquiera de los valores 1, 2 y 3. Encuentre $E[X]$ si

$$p(1) = 0{,}3 \qquad \text{y} \qquad p(2) = 0{,}5$$

6. X es una variable aleatoria que puede tomar cualquiera de los valores $1, 2, \ldots, n$, con probabilidades iguales. Esto es,

$$P\{X = i\} = \frac{1}{n} \qquad i = 1, \ldots, n$$

   (a) Si $n = 2$, encuentre $E[X]$.
   (b) Si $n = 3$, encuentre $E[X]$.
   (c) Si $n = 4$, encuentre $E[X]$.

(d) Para un n cualquiera, ¿cuál es el valor de $E[X]$?

(e) Compruebe la respuesta del apartado (d) utilizando la identidad algebraica

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$

7. Se lanza una pareja de dados. Encuentre el valor esperado del:

   (a) menor de los resultados obtenidos

   (b) mayor de los resultados obtenidos

   (Si se obtiene el mismo resultado con los dos dados, tómelo como el menor y el mayor valor.)

8. La empresa suministradora de electricidad ha informado a una compañía de *software* de ordenadores de que existe una probabilidad del 25% de que se produzca un corte de electricidad a lo largo del próximo día laborable. La compañía estima en 400 dólares el coste que deberá afrontar si sus empleados no usan sus ordenadores ese día; y, además, estima en 1200 dólares el coste si se produce el corte de electricidad cuando sus empleados están usando los ordenadores. Si la compañía pretende minimizar el valor esperado de su pérdida, ¿debería arriesgarse a utilizar los ordenadores?

9. Una compañía de ingeniería debe decidir si prepara o no una oferta para un proyecto de construcción. Preparar la oferta costará 800 dólares. Si la prepara, la compañía obtendría un beneficio bruto (excluyendo el coste de preparación) de 0 dólares si no obtiene el contrato, de 3000 dólares si obtiene el contrato y el tiempo es malo, y de 6000 dólares si obtiene el contrato y el tiempo no es malo. Si la probabilidad de conseguir el contrato es 0,4 y la de que el tiempo sea malo es 0,6, ¿cuál es el beneficio neto esperado, si la compañía prepara la oferta?

10. Toda la sangre donada a un banco sanguíneo es analizada antes de usarla. Para reducir el número total de los análisis, el banco toma pequeñas muestras de cuatro donantes distintos y las mezcla. Después se analiza la mezcla. Si se considera aceptable, el banco almacena la sangre de esas cuatro personas para un uso futuro. Si la mezcla no resulta aceptable, se analiza la sangre de cada una de las cuatro personas independientemente. Por consiguiente, por cada cuatro donantes es necesario hacer un solo análisis o bien cinco. Encuentre el número esperado de análisis que se precisan hacer si la sangre de cada donante es, independientemente, inaceptable con una probabilidad de 0,1.

11. Se elige aleatoriamente a dos personas de un grupo formado por 10 hombres y 20 mujeres. Denote por X el número de hombres elegidos y por Y el número de mujeres.

    (a) Encuentre $E[X]$.

    (b) Encuentre $E[Y]$.

    (c) Encuentre $E[X+Y]$.

12. Los dos equipos que participan en una competición a varios juegos tienen la misma probabilidad de ganar cada juego, independientemente de los resultados de los juegos anteriores. Las probabilidades de que la competición se termine en 4, 5, 6 o 7 juegos son,

respectivamente, 1/8, 1/4, 5/16 y 5/16. Bajo la hipótesis citada, ¿cuál es el número esperado de juegos de dicha competición?

13. Una compañía que regenta una cadena de establecimientos de *hardware* debe decidir en qué dirección, de dos posibles, abre un nuevo establecimiento. Si elige la primera dirección, la compañía piensa que, si el establecimiento tiene éxito, obtendrá un beneficio de 40 000 $ en el primer año y que, si el establecimiento no tiene éxito, incurrirá en una pérdida de 10 000 $. En la segunda dirección, la compañía piensa que, si se tiene éxito, durante el primer año puede tener un beneficio de 60 000 $ y, si no lo tiene, una pérdida de 25 000 $.

    (a) Si la probabilidad de éxito es de 1/2 en ambas direcciones, ¿en cuál de ellas se tendrá un mayor beneficio esperado durante el primer año?

    (b) Repita el apartado (a), asumiendo ahora que la probabilidad de que el establecimiento tenga éxito es de 1/3 en ambas direcciones.

14. Si llueve mañana, cierta persona ganará 200 dólares dando clases; si el tiempo es seco, ganará 300 dólares haciendo un trabajo de construcción. Si la probabilidad de que llueva es de 1/4, ¿cuál será la cantidad esperada que ganará mañana dicha persona?

15. Si una persona tiene una probabilidad de 1/10 de ganar 400 dólares y una probabilidad de 9/10 de ganar −50 $ (es decir, de perder 50 dólares), ¿cuál es su ganancia esperada?

16. Si una inversión presenta una probabilidad de 0,4 de obtener un beneficio de 30 000 $ y una probabilidad de 0,6 de perder 15 000 $, ¿esta inversión tiene una ganancia esperada positiva?

17. Comprobar el funcionamiento de una pieza de una máquina cuesta 40 dólares. Si se instala una pieza defectuosa, reparar el daño ocasionado a la máquina cuesta 950 dólares. Desde el punto de vista de minimizar el coste esperado, determine si se debería instalar la pieza sin comprobarla, sabiendo que la probabilidad de que sea defectuosa es de:

    (a) 0,1

    (b) 0,05

    (c) 0,01

    (d) Para que sea indiferente que se compruebe o no la pieza antes de su instalación, ¿cuál debe ser la probabilidad de que la pieza sea defectuosa?

18. Una apuesta (moralmente) correcta es aquella en la que la ganancia esperada es igual a 0. Si se apuesta una unidad monetaria en un número de la ruleta, se ganarán 35 unidades si sale dicho número y se perderá 1 unidad si no sale. Si la rueda de la ruleta está perfectamente equilibrada, la probabilidad de que salga el número apostado es de 1/38. ¿Cuál es la ganancia esperada por la apuesta de 1 unidad? ¿La apuesta es moralmente correcta?

19. Una escuela organiza una rifa en la que cada número cuesta 1 dolar. Se ofrecen 7 premios: 1 de 100 $, 2 de 50 $ y 4 de 25 $. Supongamos que se compra un número de los 500 posibles, ¿cuál es la ganancia esperada? (*Sugerencia*: La ganancia será −1, si no se gana ningún premio; 24, si se gana un premio de 25 $; 49, si se gana un premio de 50 $; o 99, si se gana un premio de 100 $.)

20. Una rueda de la ruleta consta de 18 números coloreados en rojo, 18 coloreados en negro y 2 (cero y doble cero) no coloreados. Si se apuesta 1 unidad al color rojo, uno gana 1 unidad si sale rojo o pierde 1 unidad si no sale. ¿Cuál es la ganancia esperada?

21. El primer jugador que gane 2 sets será el ganador de un partido de tenis. Supongamos que, independientemente de lo ocurrido en los sets previos, cada jugador tiene una probabilidad de 1/2 de ganar el siguiente set. Determine el número esperado de sets jugados.

22. Suponga que, en el problema 21, los dos jugadores no tienen la misma habilidad y que el jugador 1 gana cada set, independientemente de los resultados de los sets anteriores, con probabilidad 1/3.

    (a) Encuentre el número esperado de sets jugados.

    (b) ¿Cuál es la probabilidad de que el jugador 1 gane el partido?

23. Una compañía de seguros ofrece una póliza de seguro de vida que cuesta 1400 $ al año, con la que, si el asegurado fallece, la compañía se compromete a pagar 250 000 $. Si un asegurado tiene una probabilidad de fallecer de 0,005 en el transcurso del año, ¿qué beneficio esperado anual aporta este asegurado a la compañía?

24. En el ejemplo 5.8, encuentre, tanto en (a) como en (b), el valor esperado de la suma de las pagas extra de la esposa y el marido.

25. Si $E[X] = \mu$, ¿cuánto vale $E[X - \mu]$?

26. Cuatro autobuses que transportan a 148 estudiantes de una misma escuela llegan a un estadio de fútbol. Los autobuses llevan, respectivamente, a 40, 33, 50 y 25 estudiantes. Se elige aleatoriamente a uno de los estudiantes y se denota como X al número de estudiantes que iban en su autobús. También se selecciona aleatoriamente a uno de los 4 conductores y se denota como Y al número de pasajeros de su autobús.

    (a) Calcule $E[X]$ y $E[Y]$.

    (b) ¿Puede justificar intuitivamente por qué $E[X]$ es mayor que $E[Y]$?

27. El dueño de un pequeño vivero debe decidir el número de árboles de Navidad que va a almacenar. En el vivero se compran los árboles por 6 dólares y se venden por 20 dólares. Los árboles que no se venden no tienen utilidad alguna. El dueño del vivero estima que la distribución de probabilidad de la demanda de árboles es la siguiente:

Cantidad demandada	1200	1500	1800
Probabilidad	0,5	0,2	0,3

    Determine el beneficio esperado por el dueño del vivero si almacena:

    (a) 1200 árboles

    (b) 1500 árboles

    (c) 1800 árboles

28. Repita el problema 27, pero ahora suponiendo que por cada árbol no vendido se pueden recuperar 2 $.

29. La demanda diaria de una cierta tarta en una pastelería es la siguiente:

Demanda diaria	0	1	2	3	4
Probabilidad	0,15	0,25	0,30	0,15	0,15

Confeccionar cada tarta cuesta a la pastelería 4 dólares y se vende por 20 dólares. Al final del día las tartas no vendidas se tiran. ¿Tiene la pastelería un mayor beneficio esperado si confecciona 2, 3 o 4 tartas diarias?

30. Si $E[X] = 5$ y $E[Y]] = 12$, encuentre:

(a) $E[3X + 4Y]$

(b) $E[2 + 5Y + X]$

(c) $E[4 + Y]$

31. Determine la suma esperada obtenida al lanzar dos dados bien construidos:

(a) Usando la distribución de probabilidad de la suma.

(b) Aplicando un razonamiento similar al empleado en el ejemplo 5.5, junto con el hecho de que el valor esperado de una suma de variables aleatorias es igual a la suma de los valores esperados de cada variable.

32. En una pareja, la paga extra de fin de año del marido es

0	con probabilidad 0,3
1000 $	con probabilidad 0,6
2000 $	con probabilidad 0,1

La paga extra de la mujer es de:

1000 $	con probabilidad 0,7
2000 $	con probabilidad 0,3

Si S es la suma de las dos pagas extra, y encuentre $E[S]$.

33. Los datos siguientes muestran las cifras de quiebras bancarias en Estados Unidos en los años comprendidos entre 1995 y 2002.

Año	**Quiebras**
1995	8
1996	6
1997	1
1998	3
1999	8
2000	7
2001	4
2002	11

Supongamos que un comité del Congreso decide elegir aleatoriamente 2 de dichos años y analizar los expedientes de quiebra presentados en ellos. Determine el número esperado de expedientes de quiebra que deberá analizar el comité.

34. Repita el problema 33 suponiendo, en esta ocasión, que el comité elige aleatoriamente 3 de los años.

35. Una pequeña empresa de taxis dispone de 4 vehículos. A lo largo de un mes, cada taxi recibe: 0 multas de tráfico con probabilidad 0,3; 1 multa de tráfico con probabilidad 0,5, o 2 multas de tráfico con probabilidad 0,2. ¿Cuál es el número esperado de multas de tráfico que acumula la flota de los 4 taxis?

36. Suponga que se seleccionan aleatoriamente 2 baterías de un cajón que contiene 8 de buenas y 2 de defectuosas. Denote como W el número de baterías defectuosas seleccionadas.

    (a) Calcule $E[W]$ determinando primero la distribución de probabilidad de W. Si la primera batería seleccionada es defectuosa, X es igual a 1 y, en caso contrario, X es igual a 0. Igualmente, si la segunda batería elegida es defectuosa, Y será igual a 1 y, en caso contrario, es igual a 0.

    (b) Obtenga una ecuación que relacione X, Y y W.

    (c) Utilice la ecuación del apartado (b) para calcular $E[W]$.

5.4 Varianzas de las variables aleatorias

Resulta útil resumir las propiedades de una variable aleatoria por medio de un número reducido de medidas elegidas adecuadamente. Una de tales medidas es el valor esperado. Aunque el valor esperado representa la media ponderada de todos los valores posibles de la variable aleatoria, no proporciona información alguna acerca de la variación, o dispersión, de dichos valores. Por ejemplo, consideremos las variables aleatorias U, V, y W, cuyos valores y probabilidades asociadas son los siguientes:

$$U = 0 \quad \text{con una probabilidad de 1}$$

$$V = \begin{cases} -1 & \text{con probabilidad 1/2} \\ 1 & \text{con probabilidad 1/2} \end{cases}$$

$$W = \begin{cases} -10 & \text{con probabilidad 1/2} \\ 10 & \text{con probabilidad 1/2} \end{cases}$$

Aunque las tres variables aleatorias tienen el mismo valor esperado, igual a 0, existe claramente una menor dispersión en los valores de U que en los de V y, también, una menor dispersión en los valores de V que en los de W.

Dado que uno espera que la variable aleatoria tome valores alrededor de su media $E[X]$, una forma razonable de medir la variación de X es considerar en qué medida X tiende a separarse de su media. Esto es, se podría considerar $E[|X - \mu|]$, donde $\mu = E[X]$ y $|X - \mu|$ es el valor absoluto de la diferencia entre X y μ. Sin embargo, resulta más conveniente no considerar el valor absoluto sino el cuadrado de la diferencia.

Definición

Si X es una variable aleatoria con un valor esperado μ, la varianza de X, denotada por $\text{Var}(X)$, se define como

$$\text{Var}(X) = E[(X - \mu)^2]$$

Si se desarrolla $(X - \mu)^2$ se obtiene $X^2 - 2\mu X + \mu^2$ y, después se toma el valor esperado de cada término, tras unos sencillos cálculos se obtiene la siguiente fórmula para calcular $\text{Var}(X)$:

$$\text{Var}(X) = E[X^2] - \mu^2 \tag{5.1}$$

donde

$$\mu = E[X]$$

Por lo general, la mejor forma de calcular la varianza de X es si se utiliza la expresión (5.1).

Ejemplo 5.12 Calcule $\text{Var}(X)$, siendo X la variable aleatoria tal que

$$X = \begin{cases} 1 & \text{con probabilidad } p \\ 0 & \text{con probabilidad } 1 - p \end{cases}$$

Solución En el ejemplo 5.6 se vio que $E[X] = p$. Por consiguiente, si se usa la anterior fórmula de cálculo de la varianza, se tiene que

$$\text{Var}(X) = E[X^2] - p^2$$

Ahora bien,

$$X^2 = \begin{cases} 1^2 & \text{si } X = 1 \\ 0^2 & \text{si } X = 0 \end{cases}$$

Puesto que $1^2 = 1$ y $0^2 = 0$, se ve que

$$\begin{aligned} E[X^2] &= 1 \cdot P\{X = 1\} + 0 \cdot P\{X = 0\} \\ &= 1 \cdot p = p \end{aligned}$$

De donde,

$$\text{Var}(X) = p - p^2 = p(1 - p) \quad \blacksquare$$

Ejemplo 5.13 El resultado de una inversión (en unidades de 1000 $) es una variable aleatoria cuya distribución de probabilidad es

$$P\{X = -1\} = 0{,}7 \quad P\{X = 4\} = 0{,}2 \quad P\{X = 8\} = 0{,}1$$

Calcule Var(X), la varianza del resultado de la inversión.

Solución Calculemos en primer lugar el resultado medio de la inversión como sigue:

$$\begin{aligned}\mu = E[X] &= -1(0{,}7) + 4(0{,}2) + 8(0{,}1)\\ &= 0{,}9\end{aligned}$$

Esto es, el resultado esperado es de 900 dólares. Para calcular Var(X), se utilizará la fórmula

$$\text{Var}(X) = E[X^2] - \mu^2$$

Ahora bien, puesto que X^2 es igual a $(-1)^2$, 4^2 o 8^2 con unas probabilidades de 0,7, 0,2 y 0,1, respectivamente, se tiene que:

$$\begin{aligned}E[X^2] &= 1(0{,}7) + 16(0{,}2) + 64(0{,}1)\\ &= 10{,}3\end{aligned}$$

Por consiguiente,

$$\begin{aligned}\text{Var}(X) &= 10{,}3 - (0{,}9)^2\\ &= 9{,}49 \quad \blacksquare\end{aligned}$$

5.4.1 Propiedades de la varianza

Para cualquier variable aleatoria X y cualquier constante c, se puede demostrar que:

$$\text{Var}(cX) = c^2\,\text{Var}(X)$$
$$\text{Var}(X + c) = \text{Var}(X)$$

Esto es, la varianza del producto de una constante por una variable aleatoria es igual al cuadrado de la constante por la varianza de la variable aleatoria; y la varianza de la suma de una constante y una variable aleatoria es igual a la varianza de la variable aleatoria.

Aunque el valor esperado de una suma de variables aleatorias es siempre igual a la suma de los valores esperados, por lo general, este resultado no es cierto para la varianza. Por ejemplo, observe lo siguiente:

$$\begin{aligned}\text{Var}(X + X) &= \text{Var}(2X)\\ &= 2^2\,\text{Var}(X)\\ &\neq \text{Var}(X) + \text{Var}(X)\end{aligned}$$

Sin embargo, existe un caso importante en el que la varianza de una suma de variables aleatorias es igual a la suma de las varianzas; esto ocurre cuando las variables aleatorias son independientes. Antes de presentar este resultado se debe introducir el concepto de variables aleatorias independientes.

Se dice que X e Y son independientes si el conocimiento del valor de una de ellas no cambia las probabilidades de la otra. Esto es, si X toma los valores x_i, $i \geq 1$, e Y toma los valores y_j, $j \geq 1$, se dice que X e Y son independientes si los sucesos de que X sea igual a x_i y de que Y sea igual a y_j son independientes para cualquier valor x_i e y_j.

Definición

Las variables aleatorias X e Y son independientes si el conocimiento del valor de una de ellas no cambia las probabilidades de la otra.

Se verifica que la varianza de una suma de variables aleatorias independientes es igual a la suma de las varianzas.

Resultado útil

Si X e Y son variables aleatorias independientes,

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

Con mayor generalidad, si $X_1, X_2, \ldots, X_k$ son variables aleatorias independientes,

$$\text{Var}\left(\sum_{i=1}^{k} X_i\right) = \sum_{i=1}^{k} \text{Var}(X_i)$$

Ejemplo 5.14 Determine la varianza de la suma de los resultados obtenidos cuando se lanzan dos dados.

Solución Numere los dados, y sea X el valor del lanzamiento del primer dado, e Y el valor del lanzamiento del segundo. Se pide calcular $\text{Var}(X + Y)$. Puesto que los resultados de cada uno de los lanzamientos son independientes, se sabe que

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

Para calcular $\text{Var}(X)$, la varianza del resultado del primer lanzamiento, recuerde que, en el ejemplo 5.5, se vio que

$$E[X] = \frac{7}{2}$$

Puesto que X^2 puede tomar los valores 1^2, 2^2, 3^2, 4^2, 5^2 y 6^2 con probabilidades iguales, se tiene que

$$E[X^2] = \frac{1}{6}(1 + 4 + 9 + 16 + 25 + 36) = \frac{91}{6}$$

Por consiguiente,

$$\text{Var}(X) = E[X^2] - \left(\frac{7}{2}\right)^2$$

$$= \frac{91}{6} - \frac{49}{4}$$

$$= \frac{35}{12}$$

Puesto que Y tiene la misma distribución de probabilidad que X, su varianza es igualmente 35/12, de donde

$$\text{Var}(X + Y) = \frac{35}{12} + \frac{35}{12} = \frac{35}{6}$$ ■

La raíz cuadrada positiva de la varianza se denomina *desviación típica* o *estándar* (SD).

Definición

El valor SD(X), definido por

$$\text{SD}(X) = \sqrt{\text{Var}(X)}$$

se denomina *desviación típica* (o *estándar*) de X.

La desviación típica, al igual que el valor esperado, se mide en las mismas unidades que la propia variable aleatoria. Esto es, si X se mide en millas, también se medirán en millas el valor esperado y la desviación típica de X. Para calcular la desviación típica de una variable aleatoria basta con calcular su varianza y, después, extraer su raíz cuadrada.

Ejemplo 5.15 La ganancia bruta anual (en unidades de 1000 $) de un determinado cantante de rock es una variable aleatoria con un valor esperado de 400 000 $ y una desviación típica de 80 000 $. El manager del cantante recibe el 15% de la ganancia del cantante. Determine el valor esperado y la desviación típica de la cantidad recibida por el manager.

Solución Si X denota la ganancia (en unidades de 1000 $) del cantante, la ganancia del manager será de 0,15X. Su valor esperado se obtiene como sigue:

$$E[0{,}15X] = 0{,}15E[X] = 60$$

Para calcular la desviación típica, se obtendrá primero la varianza:

$$\text{Var}(0{,}15X) = (0{,}15)^2 \text{Var}(X)$$

Si se extrae la raíz cuadrada en los dos miembros de la igualdad anterior, se obtiene

$$\text{SD}(0{,}15X) = 0{,}15 \text{ SD}(X) = 12$$

Por consiguiente, la cantidad recibida por el manager es una variable aleatoria con un valor esperado de 60 000 $ y con una desviación típica de 12 000 $. ■

Problemas

1. Determine las varianzas de las variables aleatorias U, V y W, definidas al comienzo de la sección 5.4.
2. Sea $p(i) = P\{X = i\}$. Considere que

   (a) $p(0) = 0{,}50, p(1) = 0{,}50$

   (b) $p(0) = 0{,}60, p(1) = 0{,}40$

   (c) $p(0) = 0{,}90, p(1) = 0{,}10$

   ¿En qué caso piensa que Var(X) es mayor? Y ¿en cuál Var(X) es menor? Determine las varianzas reales y compruébense las contestaciones anteriores.
3. Suponga que, para alguna constante c, $P\{X = c\} = 1$. Calcule Var(X).
4. Calcule las varianzas de las variables aleatorias especificadas en el problema 1 de la sección 5.3.
5. Obtenga Var(X) para la variable aleatoria X dada en el problema 5 de la sección 5.3.
6. Si una persona tiene probabilidad 1/3 de ganar 300 $ y probabilidad 2/3 de ganar 600 $, ¿cuál es la varianza de la cantidad que dicha persona gana?
7. Encuentre la varianza del número de sets jugados en la situación descrita en el problema 21 de la sección 5.3.
8. Una pequeña empresa electrónica fundada hace 4 años tiene en la actualidad 60 empleados. A continuación se muestra la tabla de frecuencias del número de años (redondeados a enteros) que llevan en la empresa los empleados citados.

Número de años	Frecuencia
1	12
2	25
3	16
4	7

Suponga que se elige aleatoriamente a uno de los empleados y que X denota el número de años que él o ella lleva trabajando en la empresa. Encuentre

(a) $E[X]$

(b) $\text{Var}(X)$

9. La duración de las vacaciones disfrutadas por un trabajador de cierta compañía depende del resultado económico de ésta. Suponga que Fong, un empleado de esta compañía, tendrá

0 semanas de vacaciones	con probabilidad 0,4
1 semana de vacaciones	con probabilidad 0,2
2 semanas de vacaciones	con probabilidad 0,4

Suponga, también, que Fontánez, otro empleado, tendrá

0 semanas de vacaciones	con probabilidad 0,3
1 semana de vacaciones	con probabilidad 0,4
2 semanas de vacaciones	con probabilidad 0,3

X e Y denotan el número de semanas de vacaciones de Fong y de Fontanez, respectivamente.

(a) ¿Cuál piensa que es mayor, $\text{Var}(X)$ o $\text{Var}(Y)$?

(b) Calcule $\text{Var}(X)$.

(c) Calcule $\text{Var}(Y)$.

10. En el problema 27(b) de la sección 5.3, encuentre la varianza del beneficio obtenido por el vivero.

11. Se lanzan dos monedas bien construidas. Determine $\text{Var}(X)$, siendo X el número de caras resultantes.

(a) Utilice la definición de la varianza.

(b) Utililice el hecho de que la varianza de una suma de variables aleatorias independientes es igual a la suma de las varianzas.

12. Calcule la varianza del número de multas recibidas por la flota de taxis, con los datos indicados en el problema 35 de la sección 5.3. Asuma que las multas recibidas por los distintos taxis son independientes.

13. Una abogada debe decidir entre cobrar a un cliente una tarifa fija de 2000 $ o bien una tarifa condicionada de 8000 $, que sólo cobraría si gana el caso (es decir, recibiría 0 $ si lo pierde). Ella estima que la probabilidad de que gane el caso es de 0,3. Determine la desviación típica de la cantidad recibida si:

(a) Opta por la tarifa fija.

(b) Opta por la tarifa condicionada.

14. Encuentre la desviación estándar de la cantidad ganada que se especifica en el problema 14 de la sección 5.3.

15. La siguiente tabla de frecuencias muestra el número de asignaturas en las que están matriculados los 210 estudiantes de primer curso de una universidad.

Número de asignaturas	Frecuencia
1	2
2	15
3	37
4	90
5	49
6	14
7	3

Si X denota el número de asignaturas en las que está matriculado un estudiante elegido aleatoriamente, encuentre

(a) $E[X]$

(b) $\mathrm{SD}(X)$

16. El sueldo de Robert tiene un valor esperado de 30 000 $ y una desviación típica de 3000 $. El sueldo de su esposa, Sandra, tiene un valor esperado de 32 000 $ y una desviación estándar de 5000 $. Determine

(a) El valor esperado

(b) La desviación típica

del sueldo total de la familia. Para contestar al apartado (b), asuma que los sueldos de Robert y Sandra son independientes. [*Sugerencia*: Para contestar al apartado (b), calcule primero la varianza del sueldo total de la familia.]

17. Si $\mathrm{Var}(X) = 4$, ¿cuánto vale $\mathrm{SD}(3X)$? [*Sugerencia*: Calcule primero $\mathrm{Var}(3X)$.]

18. Si $\mathrm{Var}(2X + 3) = 16$, ¿cuál es el valor de $\mathrm{SD}(X)$?

19. Si X e Y son variables aleatorias independientes, ambas con varianza 1, calcule :

(a) $\mathrm{Var}(X+Y)$

(b) $\mathrm{Var}(X-Y)$

5.5 Variables aleatorias binomiales

Un tipo muy importante de variables aleatorias es el formado por las binomiales, que surgen como sigue. Supongamos que se llevan a cabo n subexperimentos (o *pruebas*) independientes, en cada uno de los cuales se puede obtener un "éxito" con una probabilidad p, o un "fracaso" con una probabilidad $1 - p$. Si X representa el número de éxitos que ocurren en las n pruebas, X se dice que es una variable aleatoria *binomial* con parámetros n y p.

Antes de obtener la fórmula general para la probabilidad de que una variable aleatoria binomial X tome cada uno de los valores posibles $0, 1, \ldots, n$, se considerará un caso particular. Supongamos que $n = 3$ y que se pretende calcular la probabilidad de que X sea igual a 2. Es decir, se pretende calcular la probabilidad de obtener exactamente 2 éxitos en tres pruebas independientes con probabilidad p de éxito, en cada una de ellas. Para obtener la probabilidad citada, consideremos todos los resultados que conducen exactamente a 2 éxitos:

$$(e, e, f),\ (e, f, e),\ (f, e, e)$$

Por ejemplo, el resultado (e,f,e) significa que en la primera prueba se obtiene un éxito; en la segunda, un fracaso; y en la tercera, otro éxito. Ahora bien, de la independencia de todas las pruebas se desprende que cada uno de estos resultados tiene una probabilidad de $p^2(1 - p)$. Por ejemplo, si E_i denota el suceso de que el resultado de la prueba i sea un éxito y F_i denota el suceso de que el resultado de la prueba i sea un fracaso, se tiene:

$$\begin{aligned} P(s, f, s) &= P(S_1 \cap F_2 \cap S_3) \\ &= P(S_1)P(F_2)P(S_3) \quad \text{por la independencia} \\ &= p(1 - p)p \end{aligned}$$

Puesto que cada uno de los tres resultados en los que aparecen exactamente 2 éxitos constan de dos éxitos y un fracaso, se puede razonar de una forma similar a la anterior que todos ellos ocurren con probabilidad $p^2(1 - p)$. Por consiguiente, la probabilidad de obtener exactamente 2 éxitos en las 3 pruebas es $3p^2(1 - p)$.

Consideremos ahora el caso general, en el que se llevan a cabo n pruebas independientes. Denotemos por X el número de éxitos obtenidos. Para determinar $P\{X = i\}$, consideremos cualquier resultado con exactamente i éxitos. Puesto que en este resultado existen i éxitos y $n - i$ fracasos, se desprende de la independencia de las pruebas que su probabilidad de ocurrencia es $p^i(1 - p)^{n-i}$. Esto es, cada uno de los resultados en los que $X = i$ tiene la misma probabilidad de ocurrencia, siendo ésta igual a $p^i(1 - p)^{n-i}$. Por consiguiente, $P\{X = i\}$ será igual a la probabilidad de ocurrencia anterior multiplicada por el número total de resultados con exactamente i éxitos. Ahora bien, se puede ver que existen $n!/[i!(n - i)!]$ resultados distintos compuestos por i éxitos y $n - i$ fracasos, donde $n!$ (léase "n factorial") es igual a 1 si $n = 0$, y es igual al producto de los n primeros números naturales en otro caso. Es decir,

$$\begin{aligned} 0! &= 1 \\ n! &= n \cdot (n - 1) \cdot \cdots 3 \cdot 2 \cdot 1 \quad \text{si } n > 0 \end{aligned}$$

Una variable aleatoria binomial con parámetros n y p representa el número de éxitos en n pruebas independientes, cuando en cada prueba se obtiene éxito con probabilidad p. Si X denota dicha variable aleatoria, para $i = 0, 1, \ldots, n$,

$$P\{X = i\} = \frac{n!}{i!(n - i)!} p^i(1 - p)^{n - i}$$

Como comprobación de la ecuación anterior, observe que en ella se establece que la probabilidad de que no ocurra ningún éxito en las n pruebas es

$$P\{X = 0\} = \frac{n!}{0!(n-0)!} p^0 (1-p)^{n-0}$$

$$= (1-p)^n \qquad \text{puesto que } 0! = p^0 = 1$$

Lo anterior es claramente correcto, ya que la probabilidad de que ocurran 0 éxitos coincide con la probabilidad de que en todas las pruebas se obtenga fracaso, y esta última probabilidad es, por la independencia, igual a $(1-p)(1-p)\cdots(1-p) = (1-p)^n$.

En la figura 5.3 se presentan las probabilidades de tres variables aleatorias binomiales con parámetros, respectivamente, $n = 10, p = 0{,}5$, $n = 10, p = 0{,}3$ y $n = 10, p = 0{,}6$.

Ejemplo 5.16 Se lanzan tres monedas bien construidas. Si los resultados de cada una son independientes, determine la probabilidad de que salgan i caras, para $i = 0, 1, 2, 3$.

Solución Si X denota el número de caras ("éxitos") obtenido, X es una variable aleatoria binomial con parámetros $n = 3, p = 0{,}5$. De lo anterior se desprende que:

$$P\{X = 0\} = \frac{3!}{0!\,3!}\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^3 = \left(\frac{1}{2}\right)^3 = \frac{1}{8}$$

$$P\{X = 1\} = \frac{3!}{1!\,2!}\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^2 = 3\left(\frac{1}{2}\right)^3 = \frac{3}{8}$$

$$P\{X = 2\} = \frac{3!}{2!\,1!}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^1 = 3\left(\frac{1}{2}\right)^3 = \frac{3}{8}$$

$$P\{X = 3\} = \frac{3!}{3!\,0!}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^0 = \left(\frac{1}{2}\right)^3 = \frac{1}{8} \quad \blacksquare$$

Ejemplo 5.17 Suponga que un cierto rasgo (tal como el color de los ojos, o el ser zurdo o no) se determina por un par de genes, y que además d representa un gen dominante, y r un gen recesivo. Una persona con una pareja de genes (d, d) se dice que es *dominante puro*, una con la pareja de genes (r, r) se dice que es *recesiva pura*, y una con la pareja de genes (d, r) se dice que es *híbrida*. En apariencia, los dominantes puros y los híbridos son similares. Los descendientes de una pareja reciben un gen de cada progenitor, y este gen puede, con la misma probabilidad, ser uno cualquiera de los dos que posee el progenitor citado.

(a) ¿Cuál es la probabilidad de que un descendiente de dos progenitores híbridos tenga la apariencia contraria (recesiva) a la de ellos?

(b) Suponga que dos padres híbridos tienen 4 descendientes. ¿Cuál es la probabilidad de que 1 de los 4 descendientes tenga una apariencia recesiva?

Solución

(a) El descendiente tendrá la apariencia recesiva si recibe los genes recesivos de cada progenitor. Por la independencia, la probabilidad de este hecho es $(1/2)(1/2) = 1/4$.

(b) Asumiendo que los genes recibidos por los descendientes son independientes (lo cual es una hipótesis usual en genética), se sigue del apartado (a) que el número de des-

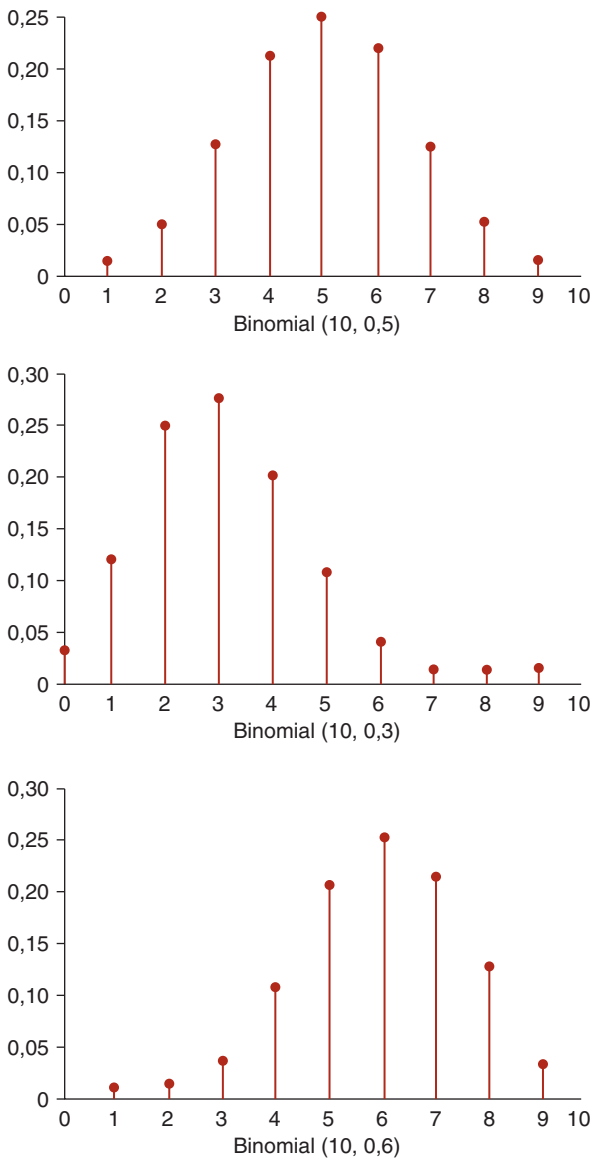


Figura 5.3 Probabilidades binomiales.

cendientes que tienen la apariencia recesiva es una variable aleatoria binomial con parámetros $n = 4$ y $p = 1/4$. Por consiguiente, si X representa el número de descendientes que tienen la apariencia recesiva, se tendrá que:

$$P\{X = 1\} = \frac{4!}{1!\,3!}\left(\frac{1}{4}\right)^1\left(\frac{3}{4}\right)^3$$

$$= 4\left(\frac{1}{4}\right)\left(\frac{3}{4}\right)^3$$

$$= \frac{27}{64}$$

Supongamos que X es una variable aleatoria binomial con parámetros n y p, y que se pretende calcular la probabilidad de que X sea menor o igual que un cierto valor j. En principio, se podría calcular dicha probabilidad como sigue:

$$P\{X \leq j\} = \sum_{i=0}^{j} P\{X = i\} = \sum_{i=0}^{j} \frac{n!}{i!(n-i)!} p^i (1-p)^{n-i}$$

Los cálculos necesarios para llevar a cabo la expresión anterior podrían ser muy costosos. Para facilitarlos, la tabla D.5 (del Apéndice D) proporciona los valores $P\{X \leq j\}$, para $n \leq 20$ y para distintos valores de p. Adicionalmente, se puede utilizar el Programa 5-1. En este programa, uno introduce los parámetros de la binomial y el valor deseado de j, y se obtiene como salida la probabilidad de que la binomial sea menor o igual que j, la probabilidad de que la binomial sea igual a j y la probabilidad de que la binomial sea mayor o igual que j. ■

Ejemplo 5.18

(a) Determine $P\{X \leq 12\}$, siendo X una variable aleatoria binomial con parámetros 20 y 0,4.

(b) Determine $P\{Y \geq 10\}$, siendo Y una variable aleatoria binomial con parámetros 16 y 0,5.

Solución A partir de la tabla D.5, se ve que:

(a) $P\{X \leq 12\} = 0{,}9790$

(b) $P\{Y \geq 10\} = 1 - P\{Y < 10\} = 1 - P\{Y \leq 9\} = 1 - 0{,}7728 = 0{,}2272$

Se podría haber ejecutado el Programa 5-1 para obtener los resultados siguientes:

La probabilidad de que una binomial (20, 0,4) sea menor o igual que 12 es de 0,978969.

La probabilidad de que una binomial (16, 0,5) sea mayor o igual que 10 es de 0,2272506. ■

5.5.1 Valor esperado y varianza de una variable aleatoria binomial

Una variable aleatoria, X, binomial (n, p) es igual al número de éxitos obtenidos en n pruebas independientes, en cada una de las cuales la probabilidad de éxito es p. En consecuencia, se puede representar X mediante la suma

$$X = \sum_{i=1}^{n} X_i$$

donde X_i es igual a 1 si en la prueba i resulta un éxito y es igual a 0 si en la prueba i resulta un fracaso. Puesto que

$$P\{X_i = 1\} = p \qquad \text{y} \qquad P\{X_i = 0\} = 1 - p$$

se desprende de los resultados de los ejemplos 5.6 y 5.12 que

$$E[X_i] = p \qquad \text{y} \qquad \text{Var}(X_i) = p(1 - p)$$

Por consiguiente, recordando que la esperanza de una suma de variables aleatorias es igual a la suma de sus esperanzas, se ve que

$$E[X] = np$$

Adicionalmente, puesto que la varianza de una suma de variables aleatorias independientes es igual a la suma de sus varianzas, se tiene

$$\text{Var}(X) = np(1 - p)$$

En resumen:

Si X es una binomial con parámetros n y p, se tiene que

$$E[X] = np$$

$$\text{Var}\,(X) = np(1 - p)$$

Ejemplo 5.19 Supongamos que cada tornillo producido tiene, independientemente, una probabilidad 0,01 de ser defectuoso. Encuentre el valor esperado y la varianza del número de tornillos defectuosos en un suministro de tamaño 1000.

Solución El número de tornillos defectuosos del suministro de tamaño 1000 es una variable aleatoria binomial con parámetros $n = 1000$, $p = 0{,}01$. Por consiguiente, el número esperado de tornillos defectuosos es

$$E[\text{número de tornillos defectuosos}] = 1000(0{,}01) = 10$$

y la varianza del varianza del número de tornillos defectuosos es

$$\text{Var}(\text{número de tornillos defectuosos}) = 1000(0{,}01)(0{,}99) = 9{,}9 \quad \blacksquare$$

Bettmann

Jacques Bernoulli

Perspectiva histórica

Las pruebas independientes con iguales probabilidades de éxito p fueron estudiadas inicialmente por el matemático suizo Jacques Bernoulli (1654-1705). En su libro *Ars Conjectandi* (El arte de la conjetura), publicado en 1713 por su sobrino Nicholas, ocho años después de su muerte, Bernoulli demostró que, si el número de pruebas era suficientemente grande, la proporción de éxitos es próxima a p con una probabilidad próxima a 1.

Jacques Bernoulli perteneció a la primera generación de la más famosa familia de matemáticos de todos los tiempos. En total existieron entre ocho y doce Bernoulli, pertenecientes a tres generaciones distintas, que hicieron contribuciones fundamentales a la Probabilidad, la Estadística y las Matemáticas, en general. Una dificultad para conocer el número exacto de Bernoulli que hubo es que varios de ellos tuvieron el mismo nombre. (Por ejemplo, dos de los hijos de Jean, el hermano de Jacques, se llamaban Jacques y Jean.) Otra dificultad estriba en el hecho de que varios de los Bernoulli fueron conocidos con nombres distintos en diferentes lugares. Nuestro Jacques (en ocasiones, escrito Jaques), por ejemplo, fue también conocido como Jakob (en ocasiones, escrito Jacob) y como James Bernoulli. Sin embargo, cualquiera que sea su número, la influencia de todos ellos y sus resultados fueron prodigiosos. ¡Como los Bach de la música, los Bernoulli de las matemáticas fueron una familia para la eternidad!

Problemas

1. Calcule (a) 4!, (b) 5!, (c) 7!
2. Encuentre

   (a) $\dfrac{8!}{3!\,5!}$ (b) $\dfrac{7!}{3!\,4!}$ (c) $\dfrac{9!}{4!\,5!}$

3. Dado que $9! = 362{,}880$, calcule 10!
4. Utilice la distribución de probabilidad de una variable aleatoria binomial con parámetros n y p para comprobar que

   $$P\{X = n\} = p^n$$

   Posteriormente, argumente por qué este resultado es cierto.
5. Si X es una variable aleatoria binomial con parámetros $n = 8$ y $p = 0{,}4$, encuentre

   (a) $P\{X = 3\}$

   (b) $P\{X = 5\}$

   (c) $P\{X = 7\}$

6. Cada rodamiento producido tiene, independientemente, una probabilidad 0,05 de ser defectuoso. Si se inspecciona una muestra de 5 de ellos, calcule la probabilidad de que:

   (a) Ninguno resulte defectuoso.

   (b) Dos o más sean defectuosos.

7. Supongamos que vamos a asistir a 6 partidos de jockey. Si cada partido tiene probabilidad 0,10 de que tenga prórroga, encuentre la probabilidad de que:

   (a) Al menos 1 de los partidos se prorrogue.

   (b) Como máximo 1 de los partidos se prorrogue.

8. Un sistema por satélite consta de 4 componentes y puede funcionar si al menos 2 de ellos operan correctamente. Si cada componente, independientemente, opera correctamente con probabilidad 0,8, ¿cuál es la probabilidad de que el sistema funcione?

9. Un canal de comunicaciones transmite los dígitos 0 y 1. Cada dígito transmitido puede ser erróneamente recibido, independientemente, con probabilidad 0,1. Suponga que se debe transmitir un mensaje muy importante de un solo dígito. Para reducir la probabilidad de error, se transmitirá la secuencia de dígitos 0 0 0 0 0 si el mensaje es 0, y se transmitirá le secuencia 1 1 1 1 1 si el mensaje es 1. El receptor del mensaje utiliza "la regla de la mayoría" para decodificarlo; esto es, el mensaje será decodificado como 0 si existen al menos 3 ceros, y será decodificado como 1 en caso contrario.

   (a) Para que el mensaje sea decodificado incorrectamente, ¿cuántos de los 5 dígitos recibidos deben ser incorrectos?

   (b) ¿Cuál es la probabilidad de que el mensaje sea decodificado incorrectamente?

10. Un examen tipo test de elección múltiple tiene tres respuestas para cada una de sus 5 preguntas. ¿Cuál es la probabilidad de que un estudiante conteste bien a 4 o más preguntas si marca las contestaciones al azar?

11. Un hombre afirma que está dotado de una percepción extrasensorial. Para comprobarlo, se realizan 8 lanzamientos de una moneda bien construida y se le pide que prediga por adelantado los resultados. Supongamos que obtiene 6 predicciones correctas. ¿Cuál sería la probabilidad de que realizara al menos este número de predicciones correctas si no estuviera dotado de la percepción extrasensorial que sostiene y simplemente hubiera hecho las predicciones al azar?

12. Cada disquete producido por una compañía es defectuoso con una probabilidad de 0,05, independientemente de un disquete a otro. La compañía vende los disquetes en paquetes de 10 y ofrece una garantía de devolución del dinero si algún disquete del paquete resulta defectuoso. Si todos los compradores ejercieran la garantía:

   (a) ¿Cuál es la probabilidad de que un paquete sea devuelto?

   (b) Si una persona compra tres paquetes, ¿cuál es la probabilidad de que exactamente devuelva uno de ellos?

13. Se lanza un dado bien construido en cuatro ocasiones. Calcule la probabilidad de que:

   (a) Salga un 6 al menos una vez.

(b) Salga un 6 exactamente una vez.

(c) Salga un 6 al menos dos veces.

14. Las estadísticas indican que el alcohol es la causa del 55% de los accidentes mortales de tráfico. Si se analizan los siguientes 3 accidentes mortales, encuentre la probabilidad de que el alcohol sea la causa en:

(a) todos ellos

(b) exactamente en 2 de ellos

(c) al menos en 1 de ellos

15. Los individuos que tienen dos genes de la anemia desarrollan esta enfermedad, mientras que los individuos que no tienen ningún gen de la anemia o tienen sólo 1 no la padecen. Si dos personas, ambas teniendo un solo gen, tienen descendencia, el hijo recibirá 2 genes de la anemia con una probabilidad de 1/4. Suponga que todos los miembros de 3 parejas tienen sólo 1 gen de la anemia y que cada una de las parejas citadas tiene un descendiente. Calcule la probabilidad de que:

(a) Ninguno de los descendientes reciba 2 genes de la anemia.

(b) Exactamente uno de los descendientes reciba 2 genes de la anemia.

(c) Exactamente dos de los descendientes reciban 2 genes de la anemia.

(d) Los tres descendientes reciban 2 genes de la anemia.

16. Si X es una variable aleatoria binomial con parámetros $n = 20$ y $p = 0{,}6$, calcule

(a) $P\{X \leq 14\}$ (b) $P\{X < 10\}$

(c) $P\{X \geq 13\}$ (d) $P\{X > 10\}$

(e) $P\{9 \leq X \leq 16\}$ (f) $P\{7 < X < 15\}$

17. Se lanza un dado bien construido en 20 ocasiones. Encuentre el valor esperado del número de veces que:

(a) Sale un 6. (b) Sale un 5 o un 6.

(c) Sale un número par. (d) Sale cualquier valor excepto el 6.

18. Encuentre las varianzas de las variables aleatorias indicadas en el problema 17.

19. La probabilidad de que una bombilla fluorescente funcione durante al menos 500 horas es 0,90. Si se tienen 8 de estas bombillas, calcule la probabilidad de que:

(a) Todas ellas funcionen al menos 500 horas.

(b) Exactamente 7 de ellas funcionen al menos 500 horas.

(c) ¿Cuál es el valor esperado del número de bombillas que funcionarán al menos 500 horas?

(d) ¿Cuál es la varianza del número de bombillas que funcionarán al menos 500 horas?

20. Si se lanza una moneda bien construida 500 veces, ¿cuál es la desviación típica del número de lanzamientos en los que saldrá cara?

21. El FBI ha publicado que el 44% de las víctimas de asesinato fallecieron por disparos de pistola. Si se seleccionan aleatoriamente 4 víctimas de asesinato, calcule:
    (a) La probabilidad de que todas hayan fallecido por disparos de pistola.
    (b) La probabilidad de que ninguna haya fallecido por la causa anterior.
    (c) La probabilidad de que al menos dos hayan fallecido por dicha causa.
    (d) El número esperado de víctimas asesinadas con pistolas.
    (e) La desviación típica del número de víctimas asesinadas con pistolas.
22. El número esperado de caras obtenidas en 10 lanzamientos de una moneda es 6. ¿Cuál es la probabilidad de que resulten 8 caras en los lanzamientos citados?
23. Si X es una variable aleatoria binomial con un valor esperado de 4 y una varianza de 2,4, calcule:
    (a) $P\{X = 0\}$
    (b) $P\{X = 12\}$
24. Si X es una variable aleatoria binomial con un valor esperado de 4,5 y una varianza de 0,45, encuentre:
    (a) $P\{X = 3\}$
    (b) $P\{X \geq 4\}$
25. Encuentre la media y la desviación típica de una variable aleatoria binomial con parámetros:
    (a) $n = 100, p = 0{,}5$ (b) $n = 100, p = 0{,}4$
    (c) $n = 100, p = 0{,}6$ (d) $n = 50, \ p = 0{,}5$
    (e) $n = 150, p = 0{,}5$ (f) $n = 200, p = 0{,}25$

*5.6 Variables aleatorias hipergeométricas

Supongamos que se seleccionan aleatoriamente n baterías de una caja que contiene N, de las cuales Np funcionan correctamente y $N(1 - p)$ son defectuosas. La variable aleatoria X, igual al número de baterías de la muestra seleccionada que funcionan correctamente, se denomina variable aleatoria hipergeométrica con parámetros n, N y p.

Se puede interpretar que el anterior experimento consiste en n pruebas, donde se considera que en la prueba i resulta un éxito si la i-ésima batería extraída funciona correctamente. Puesto que cada una de las N baterías tiene la misma probabilidad de ser la batería elegida en la extracción i, se tiene que se obtiene un éxito en la prueba i con una probabilidad $Np/N = p$. Por consiguiente, se puede concebir que X represente el número de éxitos en n pruebas, y en cada una la probabilidad de éxito es p. Lo que diferencia X de la variable aleatoria binomial es que las pruebas citadas no son independientes. Por ejemplo, supongamos que se extraen dos baterías de una caja que contiene cinco, de las cuales una funciona correctamente y las otras son defectuosas. (Esto es, $n = 2$, $N = 5$, $p = 1/5$.) La probabilidad de que la segunda batería extraída sea la que funciona correctamente es 1/5. Sin embargo, si la primera batería extraída fue la que funciona correctamente, la probabili-

dad condicionada de que la segunda sea la que funciona correctamente es 0 (puesto que, cuando se elige la segunda batería, las cuatro restantes baterías de la caja son todas defectuosas). Esto es, cuando las selecciones de las baterías se llevan a cabo sin reemplazar las que fueron elegidas anteriormente, las pruebas no son independientes y, en consecuencia, X no es una variable aleatoria binomial.

Si se utiliza el resultado de que en cada una de las n pruebas se obtiene un éxito con una probabilidad p, se puede demostrar que el número esperado de éxitos es np. Esto es,

$$E[X] = np$$

Adicionalmente, se puede demostrar que la varianza de la variable aleatoria hipergeométrica viene dada por

$$\text{Var}(X) = \frac{N - n}{N - 1} np(1 - p)$$

Así pues, mientras que el valor esperado de la variable aleatoria hipergeométrica con parámetros n, N y p coincide con el de la binomial con parámetros n, p, la varianza de la primera es menor que la de la segunda, ya que la varianza de la hipergeométrica es igual a la de la binomial multiplicada por el factor $(N - n)/(N - 1)$.

Ejemplo 5.20 Si se seleccionan aleatoriamente 6 personas de un grupo compuesto por 12 hombres y 8 mujeres, el número de mujeres elegidas es una variable aleatoria hipergeométrica con parámetros $n = 6$, $N = 20$, $p = 8/20 = 0{,}4$. Su media y su varianza son:

$$E[X] = 6(0{,}4) = 2{,}4 \qquad \text{Var}(X) = \frac{14}{19} 6(0{,}4)(0{,}6) \approx 1{,}061$$

De igual forma, el número de hombres elegidos es una variable aleatoria hipergeométrica con parámetros $n = 6$, $N = 20$, $p = 0{,}6$ ■

Supongamos ahora que N, el número de baterías de la caja, es muy grande comparado con n, el número de baterías extraídas. Por ejemplo, supongamos que se eligen aleatorimente 20 baterías de un conjunto de 10 000, en el que un 90% funciona correctamente. En este caso, con independencia de qué baterías se han seleccionado previamente, en cada nueva extracción la probabilidad de que resulte una batería que funcione correctamente es aproximadamente igual a 0,9. Por ejemplo, la primera batería seleccionada funcionará correctamente con una probabilidad de 0,9. Si la primera batería seleccionada es correcta, la que se seleccione posteriormente será igualmente correcta con una probabilidad de 8999/9999 = 0,89999; mientras que, si la primera batería seleccionada resultó ser defectuosa, la probabilidad de que la batería extraída en segundo lugar funcione correctamente es de 9000/9999 = 0,90009. Unas cifras similares se pueden obtener para las restantes extracciones; así pues, se puede concluir que, cuando N es grande en relación a n, las n pruebas son aproximadamente independientes, lo que significa que X se aproxima a la variable aleatoria binomial.

Cuando N es grande con relación a n, la variable aleatoria hipergeométrica de parámetros n, N y p aproximadamente sigue una distribución binomial con parámetros n y p.

Problemas

En los problemas siguientes, diga si la variable aleatoria X es binomial o hipergeométrica. Igualmente, obtenga sus parámetros (n y p si se trata de la binomial; o n, N y p si se trata de la hipergeométrica).

1. Un lote de 200 elementos contiene 18 que son defectuosos. Sea X el número de elementos defectuosos de una muestra de 20 elementos.
2. Un restaurante sabe por experiencia que el 15% de las reservas resultan fallidas. Se han producido 20 reservas para esta noche. Sea X el número de reservas en las que los clientes no fallan.
3. En una versión del juego de la lotería, cada jugador selecciona seis números del 1 al 54. Los organizadores seleccionan igualmente 6 números de los 54, y estos últimos se convierten en los números ganadores. Denote por X la cantidad de números seleccionados por un determinado jugador que coinciden con los números ganadores.
4. Cada fusible nuevo producido es, independientemente de los restantes, defectuoso con probabilidad 0,05. Denote por X el número de fusibles defectuosos de los últimos 100 producidos.
5. Suponga que, en un conjunto de 100 fusibles, hay 5 que son defectuosos. Denote por X el número de fusibles defectuosos descubiertos cuando se seleccionan aleatoriamente 20 de ellos y se comprueban.
6. Las cartas de una baraja, primero, se barajan y, después, se van colocando sucesivamente boca arriba. Denote por X el número de ases que aparecen en las 10 primeras cartas descubiertas.
7. Se barajan las cartas de una baraja y, después, se descubre la carta de arriba. Posteriormente, esta carta se devuelve a la baraja y se repite la operación. Esto continúa hasta que se hayan descubierto 10 cartas. Denote por X el número de ases que han aparecido en las cartas descubiertas.

*5.7 Variables aleatorias de Poisson

Una variable aleatoria X que puede tomar los valores 0, 1, 2, . . . se dice que es una variable aleatoria de Poisson con un parámetro λ, si para algún valor positivo de λ sus probabilidades vienen dadas por

$$P\{X = i\} = c\lambda^i/i!, \qquad i = 0, 1, \cdots$$

En la expresión anterior, c es una constante que depende de λ. Su valor explícito es $c = e^{-\lambda}$, donde e es la conocida constante matemática cuyo valor es aproximadamente igual a 2,718.

Se dice que una variable aleatoria X es una variable aleatoria de Poisson con un parámetro λ si

$$P\{X = i\} = \frac{e^{-\lambda}\lambda^i}{i!}, \qquad i = 0, 1, \cdots$$

En la figura 5.4 se presenta el gráfico de las probabilidades de la variable aleatoria de Poisson con un parámetro $\lambda = 4$.

Ejemplo 5.21 Si X es una variable aleatoria de Poisson con un parámetro $\lambda = 2$, calcule $P\{X = 0\}$.

Solución

$$P\{X = 0\} = \frac{e^{-2}2^0}{0!}$$

Recordando que $2^0 = 1$ y $0! = 1$, se obtiene

$$P\{X = 0\} = e^{-2} = 0{,}1353.$$

En lo anterior, el valor de e^{-2} se ha obtenido de la tabla de exponenciales. Alternativamente, se podría haber obtenido con una calculadora científica de mano o con un ordenador personal. ■

Las variables aleatorias de Poisson surgen como una aproximación de las variables aleatorias binomiales. Considere n pruebas independientes, en cada una de las cuales se puede obtener un éxito con una probabilidad p o un fracaso con una probabilidad $1 - p$. Si el número de pruebas es grande y la probabilidad de éxito en cada una de ellas es pequeña, el número total de éxitos será aproximadamente una variable aleatoria de Poisson con un parámetro $\lambda = np$.

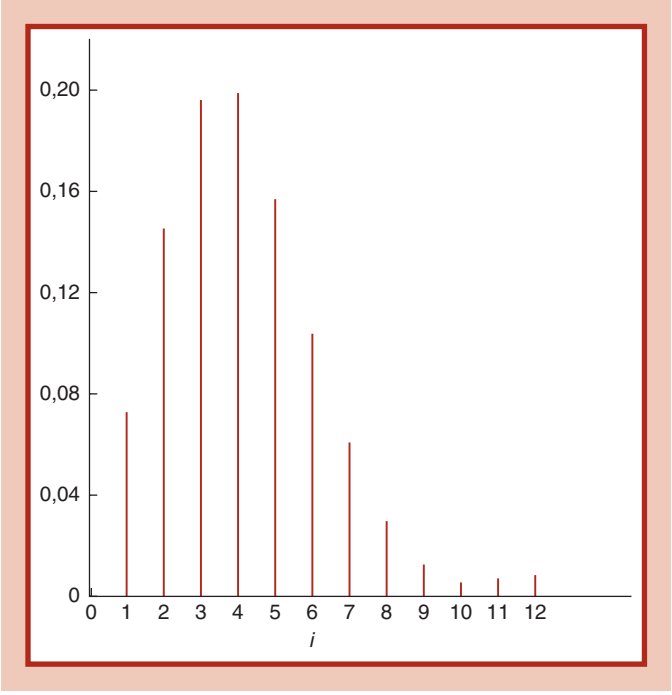


Figura 5.4 Probabilidades de una variable aleatoria de Poisson con un parámetro $\lambda = 4$.

A continuación se indican algunos ejemplos de variables aleatorias cuyas probabilidades son aproximadamente iguales que las probabilidades de la Poisson, para algún λ.

1. El número de errores de impresión en una página de un libro
2. El número de personas de una comunidad que tienen al menos 100 años
3. El número de personas que entran en una oficina de correos en un determinado día.

Cada una de ellas es aproximadamente una variable aleatoria de Poisson debido a la aproximación citada de la binomial. Se puede suponer que, en cada letra tecleada de una página, la probabilidad de cometer un error es constante; por consiguiente, el número de errores cometidos en dicha página es aproximadamente una variable aleatoria de Poisson con un parámetro $\lambda = np$, donde n es el número (grande) de letras de la página y p es la probabilidad (pequeña) de que se cometa un error en cualquier letra dada.

Ejemplo 5.22 Supongamos que las piezas producidas por una máquina son, independientemente, defectuosas con probabilidad 0,1. ¿Cuál es la probabilidad de que una muestra de 10 piezas contenga como máximo una que sea defectuosa? ¿Cuál es la aproximación de Poisson para esta probabilidad?

Solución Si X denota el número de piezas defectuosas, se tiene que X es una variable aleatoria binomial de parámetros $n = 10$, $p = 0{,}1$. Así pues, la probabilidad pedida es

$$P\{X = 0\} + P\{X = 1\} = \binom{10}{0}(0{,}1)^0(0{,}9)^{10} + \binom{10}{1}(0{,}1)^1(0{,}9)^9$$

$$= 0{,}7361$$

Puesto que $np = 10(0{,}1) = 1$, la aproximación de Poisson conduce al valor

$$P\{X = 0\} + P\{X = 1\} = e^{-1} + e^{-1} = 0{,}7358$$

Así pues, incluso en este caso, en el que n es igual a 10 (que no es muy grande) y p es igual a 0,1 (que no es muy pequeño), la aproximación de Poisson a la probabilidad binomial es bastante precisa. ■

Tanto el valor esperado como la varianza de la variable aleatoria de Poisson son iguales a λ. Esto es, se cumple lo siguiente:

Si X es una variable aleatoria de Poisson con un parámetro λ, $\lambda > 0$, se tiene que:

$$E[X] = \lambda$$

$$\text{Var}(X) = \lambda$$

Ejemplo 5.23 Supongamos que el número medio de los accidentes que ocurren semanalmente en una determinada autopista es igual a 1,2. Aproxime la probabilidad de que ocurra al menos un accidente durante la semana en curso.

Solución Denotemos como X el número de accidentes. Puesto que es razonable suponer que existe un gran número de coches que pasan por la autopista, y cada uno tiene una probabilidad muy pequeña de sufrir un accidente, el número de accidentes será aproximadamente una variable aleatoria de Poisson. Esto es, si X denota el número de accidentes que ocurrirán durante la semana, X es aproximadamente una variable aleatoria de Poisson con un valor medio de $\lambda = 1{,}2$. La probabilidad pedida se obtiene como sigue:

$$\begin{aligned} P\{X > 0\} &= 1 - P\{X = 0\} \\ &= 1 - \frac{e^{-1,2}(1,2)^0}{0!} \\ &= 1 - e^{-1,2} \\ &= 1 - 0{,}3012 \\ &= 0{,}6988 \end{aligned}$$

En consecuencia, la probabilidad de que se produzca al menos un accidente durante esta semana es aproximadamente igual al 70%. ■

Problemas

Los datos siguientes serán útiles en los problemas siguientes. Los valores dados son los correctos hasta la cuarta cifra decimal.

$$e^{-1/2} = 0{,}6065, \quad e^{-4} = 0{,}0183, \quad e^{-1} = 0{,}3679, \quad e^{-0,3} = 0{,}7408$$

1. Si X es de una variable aleatoria de Poisson con media 4, calcule:
   (a) $P\{X = 1\}$
   (b) $P\{X = 2\}$
   (c) $P\{X > 2\}$
2. Compare la aproximación de Poisson con la probabilidad binomial verdadera en los casos siguientes:
   (a) $P\{X = 2\}$ cuando $n = 10, p = 0{,}1$
   (b) $P\{X = 2\}$ cuando $n = 10, p = 0{,}05$
   (c) $P\{X = 2\}$ cuando $n = 10, p = 0{,}01$
   (d) $P\{X = 2\}$ cuando $n = 10, p = 0{,}3$

3. Una persona compra el mismo billete de lotería en 500 sorteos distintos. Si en cada sorteo la probabilidad de que gane el premio es de 1/1000, ¿cuál es la probabilidad aproximada de lo siguiente?
   (a) De que gane 0 premios.
   (b) De que gane exactamente 1 premio.
   (c) De que gane al menos 2 premios.
4. Si X es una Poisson con media $\lambda = 144$, calcule
   (a) $E[X]$ (b) $\mathrm{SD}(X)$
5. Una determinada compañía de seguros paga una media de 4 enfermedades costosas mensualmente.
   (a) Aproxime la probabilidad de que no pague ninguna enfermedad costosa el mes próximo.
   (b) Aproxime la probabilidad de que pague como máximo 2 enfermedades costosas el próximo mes.
   (c) Aproxime la probabilidad de que pague al menos 4 enfermedades costosas el próximo mes.

Términos clave

Variable aleatoria: Una magnitud cuyo valor viene determinado por el resultado de un experimento probabilístico.

Variable aleatoria discreta: Una variable aleatoria cuyos posibles valores son una sucesión de puntos distintos de la recta real.

Valor esperado de una variable aleatoria: Media ponderada de todos los posibles valores de la variable aleatoria; el peso dado a cada valor coincide con la probabilidad de que la variable aleatoria tome el valor citado. También recibe el nombre de **esperanza** o **media** de la variable aleatoria.

Varianza de una variable aleatoria: Valor esperado de las diferencias cuadráticas entre la variable aleatoria y su valor esperado.

Desviación típica de una variable aleatoria: Raíz cuadrada de la varianza.

Variables aleatorias independientes: Conjunto de variables aleatorias que tienen la propiedad de que, cuando se conocen los valores de cualquier subconjunto de ellas, este hecho no afecta a las probabilidades de las restantes variables.

Variable aleatoria binomial con parámetros n y p: Una variable aleatoria igual al número de éxitos en n pruebas independientes, cuando la probabilidad de éxito en cada prueba es igual a p.

Resumen

Una *variable aleatoria* es una magnitud cuyo valor viene determinado por el resultado de un experimento probabilístico. Si se pueden escribir sus posibles valores distintos en forma de sucesión, la variable aleatoria se denomina *discreta*.

Sea X una variable aleatoria cuyos valores posibles son x_i, $i = 1, \ldots, n$; y supongamos que X toma el valor x_i con probabilidad $P\{X = x_i\}$. El *valor esperado* de X, también conocido como *media* de X o *esperanza* de X, se denota por $E[X]$ y se define como

$$E[X] = \sum_{i=1}^{n} x_i P\{X = x_i\}$$

Si X es una variable aleatoria y c es una constante, se verifica que:

$$E[cX] = cE[X]$$

$$E[X + c] = E[X] + c$$

Para cualesquiera variables aleatorias $X_1, \ldots, X_k$,

$$E[X_1 + X_2 + \ldots + X_k] = E[X_1] + E[X_2] + \ldots + E[X_k]$$

Las variables aleatorias X e Y son *independientes* si el conocimiento del valor que toma una de ellas no modifica las probabilidades de la otra.

La *varianza* de una variable aleatoria mide la media de las distancias al cuadrado de la variable aleatoria a su media. Específicamente, si X tiene una media $\mu = E[X]$, la varianza de X, denotada por $\mathrm{Var}(X)$, se define como

$$\mathrm{Var}(X) = E[(X - \mu)^2]$$

Una propiedad de la varianza es que para cualquier constante c y variable aleatoria X, se verifica que:

$$\mathrm{Var}(cX) = c^2\mathrm{Var}(X)$$

$$\mathrm{Var}(X + c) = \mathrm{Var}(X)$$

Aunque por lo general la varianza de una suma de variables aleatorias no es igual a la suma de sus varianzas, esto sí que es cierto cuando las variables aleatorias son independientes. Es decir,

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$

si X e Y son independientes.

La raíz cuadrada de la varianza se denomina *desviación típica* (o *estándar*) y se denota como $\mathrm{SD}(X)$. Esto es,

$$\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}$$

Considere n pruebas independientes en cada una de las cuales la probabilidad de éxito es p. Si X denota el número total de éxitos en las n pruebas, se dice que X es una variable aleatoria *binomial* con parámetros n y p. Sus probabilidades vienen dadas por

$$P\{X = i\} = \frac{n!}{i!(n - i)!} p^i (1 - p)^{n-i} \quad i = 0, \ldots, n$$

En lo anterior, $n!$ (llamado *n factorial*) se define por

$$0! = 1 \quad n! = n(n-1)\ldots 3 \cdot 2 \cdot 1$$

La media y la varianza de una variable aleatoria binomial con parámetros n y p son

$$E[X] = np \quad \text{y} \quad \text{Var}(X) = np(1-p)$$

Una variable aleatoria binomial con un valor grande de n y un valor pequeño de p se puede aproximar por una variable aleatoria de Poisson, cuyas probabilidades vienen dadas por

$$P\{X = i\} = e^{-\lambda}\frac{\lambda^i}{i!}, \quad i = 0, 1, \ldots$$

donde $\lambda = np$. Tanto la media como la varianza de esta variable aleatoria son iguales a λ.

Problemas de repaso

1. Si $P\{X \leq 4\} = 0{,}8$ y $P\{X = 4\} = 0{,}2$, calcule
   (a) $P\{X \geq 4\}$
   (b) $P\{X < 4\}$
2. Si $P\{X \leq 6\} = 0{,}7$ y $P\{X < 6\} = 0{,}5$, calcule
   (a) $P\{X = 6\}$
   (b) $P\{X > 6\}$
3. Un estudiante de derecho que está a punto de graduarse no sabe si desea ejercer como abogado o dedicarse a los negocios familiares. Ha decidido que su decisión dependa de si suspende o no un examen muy importante, y hará como máximo 4 intentos. Se dedicará a la abogacía si aprueba el examen y se dedicará a los negocios familiares si suspende el examen en los cuatro intentos. Supongamos que cada vez que se presenta al examen la probabilidad de que apruebe es 0,3, con independencia de los resultados previos. Denote por X el número de veces que se presenta al examen.
   (a) ¿Cuáles son los valores posibles de X?
   (b) ¿Cuál es la distribución de probabilidad de X?
   (c) ¿Cuál es la probabilidad de que apruebe el examen?
   (d) Encuentre $E[X]$.
   (e) Calcule $\text{Var}(X)$.
4. Suponga que X puede tomar los valores 1 o 2. Si $E[X] = 1{,}6$, calcule $P\{X = 1\}$.
5. En un libro de juegos de azar se recomienda la siguiente "estrategia ganadora" en el juego de la ruleta. Un jugador debe apostar 1 unidad monetaria al rojo. Si sale rojo (lo que tiene una probabilidad 18/38 de ocurrir), el jugador recoge su ganancia de 1 unidad

y se va. Si el jugador pierde su apuesta (lo que ocurre con probabilidad 20/38), vuelve a apostar al rojo 2 unidades monetarias, tras lo cual se retira. Represéntese por X la ganancia final del jugador.

(a) Calcule $P\{X > 0\}$.

(b) ¿Está usted convencido de que la estrategia descrita es realmente "ganadora"?

(c) Calcule $E[X]$.

6. Dos personas se han de encontrar en un parque. Cada persona puede llegar con igual probabilidad e independientemente de la otra a las 3:00 h, a las 4:00 h o a las 5:00 h de la tarde. Sea X el tiempo que esperará la primera persona que llegue, siendo X igual a 0 si ambas personas llegan al mismo tiempo. Calcule $E[X]$.

7. La probabilidad de que un vendedor de coches usados venda un coche al siguiente cliente que se presente es de 0,3. Si lo vende, el coche vendido puede costar 4000 $ o 6000 $ con igual probabilidad. Denote por X la cantidad gastada por el cliente.

(a) Calcule la distribución de probabilidad de X.

(b) Encuentre $E[X]$.

(c) Encuentre $\text{Var}(X)$.

(d) Calcule $\text{SD}(X)$.

8. Se seleccionan aleatoriamente dos baterías de una caja que contiene 12, de las cuales 8 son buenas y 4 defectuosas. ¿Cuál es el número esperado de baterías defectuosas seleccionadas?

9. Una compañía está preparando una oferta para un contrato de suministro de cuadernos a varias escuelas de la ciudad. El coste del suministro de este material para la compañía es de 140 000 $. Está estudiando dos posibles ofertas alternativas: una al alza (un 25% por encima del coste) o una a la baja (un 10% por encima del coste). Por la experiencia anterior, la compañía sabe que si oferta al alza, la probabilidad de ganar el contrato es de 0,15, mientras que si oferta a la baja, la probabilidad citada es de 0,40. ¿Qué oferta maximiza el beneficio esperado de la compañía?

10. Si $E[3X + 10] = 70$, ¿cuál es el valor de $E[X]$?

11. La probabilidad de que una vendedora de aspiradoras no venda ninguna en el día de hoy es de 1/3, la probabilidad de que venda 1 es de 1/2 y la probabilidad de que venda 2 es de 1/6. Cada venta tiene, independientemente, la misma probabilidad de que se trate de una aspiradora estándar, que cuesta 500 $, o de una aspiradora de lujo, que cuesta 1000 $. Denote por X el valor total de las ventas realizadas.

(a) Encuentre $P\{X = 0\}$. (b) Encuentre $P\{X = 500\}$.

(c) Encuentre $P\{X = 1000\}$. (d) Encuentre $P\{X = 1500\}$.

(e) Encuentre $P\{X = 2000\}$. (f) Encuentre $E[X]$.

(g) Suponga que la vendedora recibe una comisión del 20% de las ventas y denote por Y la cantidad recibida. Calcule $E[X]$.

12. Las 5 familias que viven en un bloque de pisos tienen 12 hijos. Una familia tiene 4 hijos; otra tiene 3; otras dos tienen 2, y una más tiene 1. Denote por X el número de hijos de una familia elegida aleatoriamente, y por Y el número de hijos de la familia de un hijo elegido aleatoriamente. Esto es, X se refiere al experimento en el que todas las familias tienen igual probabilidad de ser seleccionadas, mientras que Y se refiere al experimento en el que cada uno de los 12 hijos tiene la misma probabilidad de ser seleccionado.

    (a) ¿Qué variable aleatoria cree que tiene mayor valor esperado, X o Y?

    (b) Calcule $E[X]$ y $E[Y]$.

13. Un agente financiero está evaluando dos posibilidades de inversión. La inversión A se traducirá en un beneficio de:

    200 000 $ con probabilidad 1/4

    100 000 $ con probabilidad 1/4

    150 000 $ con probabilidad 1/2

    La inversión B producirá un beneficio de:

    300 000 $ con probabilidad 1/8

    200 000 $ con probabilidad 1/4

    150 000 $ con probabilidad 3/8

    400 000 $ con probabilidad 1/4

    (a) ¿Cuál es el beneficio esperado de la inversión A?

    (b) ¿Cuál es el beneficio esperado de la inversión B?

    (c) Si se decide invertir tanto en A como en B, ¿cuál será el beneficio esperado?

14. Si $\text{Var}(X) = 4$, encuentre

    (a) $\text{Var}(2X + 14)$

    (b) $\text{SD}(2X)$

    (c) $\text{SD}(2X + 14)$

15. Suponga que $E[X] = \mu$ y $\text{SD}(X) = \sigma$. Sea

$$Y = \frac{X - \mu}{\sigma}$$

    (a) Demuestre que $E[Y] = 0$.

    (b) Demuestre que $\text{Var}(Y) = 1$.

    La variable aleatoria Y se denomina versión estandarizada de X. Esto es, dada una variable aleatoria, si se la resta su valor esperado y se divide el resultado por su desviación típica, la variable aleatoria resultante se dice que está estandarizada. La variable estandarizada tiene valor esperado 0 y varianza 1.

16. Un manager tiene dos clientes. La renta anual bruta del primer cliente es una variable aleatoria con valor esperado 200 000 $ y desviación típica 60 000 $. La renta anual bruta del segundo cliente es una variable aleatoria con valor esperado 140 000 $ y desviación típica 50 000 $. Si la tarifa del manager es un 15% de la renta bruta del primer cliente y un 20% de la del segundo, calcule:

    (a) el valor esperado de las ganancias del manager,

    (b) la desviación típica de la ganancia total del manager.

    Asuma en el apartado (b) que las rentas de los dos clientes son independientes.

17. Se lanza n veces una moneda truncada, cuya probabilidad de cara es 0,6. Calcule la probabilidad de que el número total de caras obtenidas en dichos lanzamientos sobrepase al número total de cruces, siendo

    (a) $n = 1$ (b) $n = 3$

    (c) $n = 5$ (d) $n = 7$

    (e) $n = 9$ (f) $n = 19$

18. Cada cliente que entra en una tienda de televisores compra una televisión de tamaño normal con probabilidad 0,3, compra un televisor de tamaño grande con probabilidad 0,1, o no compra ningún televisor con probabilidad 0,6. Calcule la probabilidad de que los próximos 5 clientes

    (a) Compren un total de 3 televisores de tamaño normal.

    (b) No compren ningún televisor de tamaño grande.

    (c) Compren un total de 2 televisores.

19. Una vendedora tiene un 60% de probabilidad de realizar una venta cada vez que visita una tienda de ordenadores. Si visita 3 tiendas de ordenadores al mes y los resultados de cada visita son independientes,

    (a) ¿Cuál es la probabilidad de que no consiga hacer ninguna venta el mes próximo?

    (b) ¿Cuál es la probabilidad de que realice 2 ventas el mes próximo?

    (c) ¿Cuál es la probabilidad de que haga al menos 1 venta en cada uno de los próximos tres meses?

20. Sea X una variable aleatoria binomial tal que

$$E[X] = 6 \quad \text{y} \quad \text{Var}(X) = 2{,}4$$

Calcule:

(a) $P\{X > 2\}$

(b) $P\{X \leq 9\}$

(c) $P\{X = 12\}$

6

Variables aleatorias normales

Entre otras peculiaridades del siglo XIX se encuentra ésta: mediante el inicio de una recogida sistemática de datos estadísticos se hizo posible el estudio de las ciencias sociales.

Alfred North Whitehead

Se introducen las variables aleatorias continuas, aquellas que pueden tomar cualquier valor dentro de un intervalo. Se ve cómo se pueden determinar sus probabilidades a partir de una curva asociada a dichas variables, conocida como *función de densidad de probabilidad*. Se estudia una clase especial de variables aleatorias continuas, compuesta por las *variables aleatorias normales*. Se introduce la variable aleatoria normal estándar y se presenta una tabla que nos permite calcular sus probabilidades asociadas. Se muestra cómo cualquier variable aleatoria normal se puede transformar en una estándar, lo que nos permite determinar sus probabilidades. Se presenta la propiedad aditiva de las variables aleatorias normales. Finalmente, se estudian los percentiles de las variables aleatorias normales.

6.1 Introducción

En este capítulo se introduce y se estudia la distribución normal. Tanto desde un punto de vista teórico como práctico, esta distribución es sin duda la más importante dentro de la Estadística.

La distribución normal pertenece a una clase de distribuciones conocida como la clase de las distribuciones *continuas*. Éstas se introducen en la sección 6.2. En la sección 6.3 se define lo que se conoce como distribución normal, y se presenta una regla de aproximación relativa a sus probabilidades. En la sección 6.4 se considera la distribución normal estándar, que es una distribución normal con media 0 y varianza 1, y se muestra cómo se determinan sus probabilidades mediante el uso de una tabla. En la sección 6.5, se ve cómo cualquier variable aleatoria normal se puede transformar en una normal estándar, y se utiliza esta transformación para determinar las probabilidades de las normales generales. La propiedad aditiva de las variables aleatorias normales se estudia en la sección 6.6. Por último, en la sección 6.7, se consideran los percentiles de estas últimas variables.

La distribución normal fue introducida en 1733 por el matemático francés Abraham De Moivre.

6.2 Variables aleatorias continuas

Mientras que los valores posibles de una variable aleatoria discreta se pueden escribir como una sucesión de puntos aislados, una *variable aleatoria continua* es aquella cuyo conjunto de valores posibles es un intervalo. Es decir, una variable aleatoria continua puede tomar cualquier valor comprendido dentro de cierto intervalo. Por ejemplo, variables tales como el tiempo que se tarda en llevar a cabo un determinado experimento científico o el peso de un individuo se considera que son variables aleatorias continuas.

Toda variable aleatoria continua X tiene una curva asociada a ella. Se puede utilizar esta curva, formalmente conocida como la *función de densidad de probabilidad* de la variable, para obtener las probabilidades referidas a X. Esto se puede llevar a cabo como sigue. Considere dos puntos cualesquiera a y b, siendo a menor que b. La probabilidad de que X tome un valor comprendido entre a y b es igual al área bajo la curva dentro de este intervalo. Esto es,

$$P\{a \leq X \leq b\} = \text{área bajo la curva entre } a \text{ y } b$$

La figura 6.1 muestra una función de densidad de probabilidad.

Puesto que X debe asumir algún valor, se tiene que el área total bajo la curva de densidad debe ser igual a 1. Adicionalmente, puesto que el área bajo la gráfica de la función de densidad de probabilidad entre los puntos a y b es la misma con independencia de que los extremos a y b se incluyan o no, se ve que

$$P\{a \leq X \leq b\} = P\{a < X < b\}$$


(North Wind Picture Archives)
Abraham De Moivre

Perspectiva histórica

Abraham De Moivre (1667–1754)

En la actualidad existe un gran número de consultores estadísticos, y algunos ejercen su oficio en las oficinas más elegantes. Sin embargo, en los primeros años del siglo XVIII, el primero de ellos trabajó en los aledaños de una oscura y destartalada casa de apuestas en Long Acres, Londres, conocida como la Slaughter´s Coffee House. Abraham De Moivre era un refugiado protestante de la Francia católica, que por un cierto precio calculaba las probabilidades de las apuestas hechas en todo tipo de juegos de azar.

A pesar de que el descubridor de la curva normal, De Moivre, se ganaba su vida en la casa de apuestas, fue un matemático reconocido en su época. De hecho, fue miembro de la Royal Society y se sabe que era íntimo amigo de Isaac Newton.

Así es como Karl Pearson imaginó a De Moivre trabajando en la casa de apuestas citada:

> *"Me imagino a De Moivre trabajando en una sucia mesa de la casa de apuestas, con un decaído apostante a su lado e Isaac Newton abriéndose paso entre la multitud para sacarle de allí. Sería una gran imagen para un artista inspirado."*

Esto es, la probabilidad de que una variable aleatoria continua caiga dentro de un intervalo es la misma independientemente de que se incluyan o no los extremos del intervalo.

La curva de densidad de probabilidad de una variable aleatoria X nunca está por debajo del eje x y tiene la propiedad de que el área total entre la curva y el eje x es igual a 1. La curva determina las probabilidades asociadas a X, de forma que el área bajo la curva entre los puntos a y b es igual a la probabilidad de que X esté entre a y b.

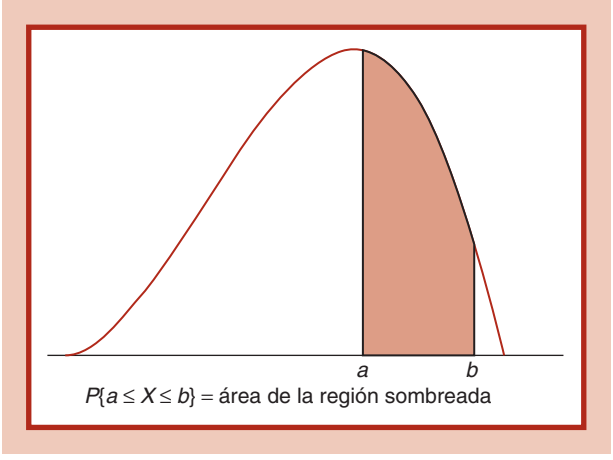


Figura 6.1 Función de densidad de probabilidad de X.

Problemas

1. La figura 6.2 es la función de densidad de probabilidad de una variable aleatoria que representa el tiempo (en minutos) que tarda un técnico en reparar un televisor. Los números en cada región indican las áreas de éstas. Calcule cuál es la probabilidad de que el técnico tarde en reparar un televisor:

   (a) menos de 20 minutos

   (b) menos de 40 minutos

   (c) más de 50 minutos

   (d) entre 40 y 70 minutos

2. Una variable aleatoria se dice que es *uniforme* en el intervalo (a, b) si el conjunto de sus valores posibles coincide con este intervalo y la gráfica de su función de densidad es horizontal. Esto es, la función de densidad coincide con la mostrada en la figura 6.3.

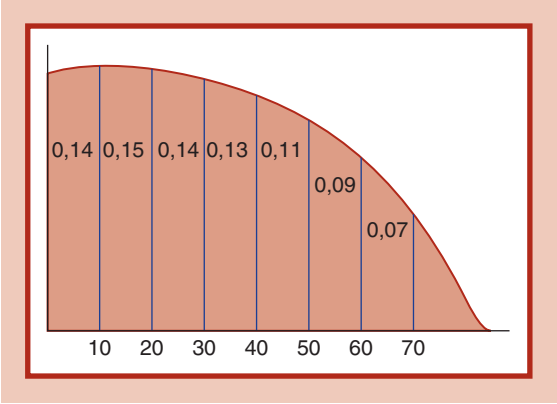


Figura 6.2 Función de densidad de probabilidad de X.

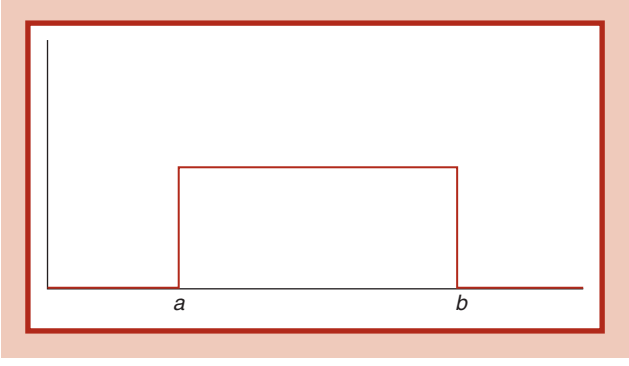


Figura 6.3 Función de densidad de la variable aleatoria uniforme (a, b).

(a) Explique por qué la altura de la gráfica de la función de densidad es $1/(b - a)$. (*Sugerencia*: Recuerde que el área por debajo la función de densidad debe ser igual a 1, y además recuerde la fórmula del área de un rectángulo.)

(b) ¿Cuál es el valor de $P\{X \leq (a + b)/2\}$?

3. Suponga que X es una variable aleatoria uniforme en el intervalo (0, 1). Esto es, $a = 0$ y $b = 1$ para la variable especificada en el problema 3. Calcule:

(a) $P\{X > 1/3\}$

(b) $P\{X \leq 0{,}7\}$

(c) $P\{0{,}3 < X \leq 0{,}9\}$

(d) $P\{0{,}2 \leq X < 0{,}8\}$

4. Una persona se ha de encontrar con un amigo a las 2 de la tarde. Aunque dicha persona es siempre puntual, su amigo siempre se retrasa y, en realidad, llegará a la cita a una hora distribuida uniformemente entre las 2 y las 3 de la tarde. Calcule la probabilidad de que la persona citada tenga que esperar:

(a) al menos 30 minutos

(b) menos de 15 minutos

(c) entre 10 y 35 minutos

(d) menos de 45 minutos

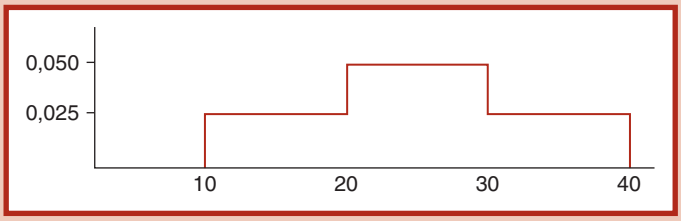
5. Suponga en el problema 4 que el amigo llega a una hora distribuida uniformemente entre las 1:30 h y las 3:00 h de la tarde. Calcule la probabilidad de que:

(a) La persona citada sea la primera en llegar.

(b) Su amigo tenga que esperar más de 15 minutos.

(c) La persona citada tenga que esperar más de 30 minutos.

6. Suponga que el número de minutos que juega un jugador de baloncesto universitario en un partido elegido aleatoriamente sigue la siguiente curva de densidad:



Calcule la probabilidad de que el jugador juegue:

(a) más de 20 minutos

(b) menos de 25 minutos

(c) entre 15 y 35 minutos

(d) más de 35 minutos

7. Denote por X el número de minutos jugados por el jugador de baloncesto del problema 6. Calcule:

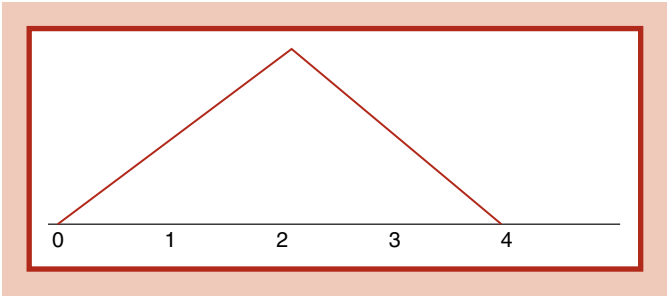
(a) $P\{20 < X < 30\}$

(b) $P\{X > 50\}$

(c) $P\{20 < X < 40\}$

(d) $P\{15 < X < 25\}$

8. Ahora son las 2 de la tarde y Elena ha planeado estudiar su examen de Estadística hasta las 6, hora en que saldrá a cenar. Sin embargo, ella sabe que tendrá interrupciones durante el estudio y piensa que el tiempo real dedicado a estudiar en las próximas 4 horas es una variable aleatoria cuya función de densidad de probabilidad es la siguiente:



(a) ¿Cuál es la altura de la curva en el valor 2? (*Sugerencia*: Recuerde la fórmula del área del triángulo.)

(b) ¿Cuál es la probabilidad de que estudie más de 3 horas?

(c) ¿Cuál es la probabilidad de que estudie entre 1 y 3 horas?

6.3 Variables aleatorias normales

El tipo más importante de variables aleatorias es la variable aleatoria normal. La función de densidad de probabilidad de una variable aleatoria normal X viene determinada por dos parámetros: el valor esperado y la desviación típica de X. Representemos estos valores por μ y σ, respectivamente. Esto es, sean

$$\mu = E[X] \qquad \text{y} \qquad \sigma = \text{SD}(X)$$

La densidad de probabilidad normal tiene una forma acampanada que es simétrica respecto del valor μ. Su variabilidad viene medida por σ. Cuanto mayor es σ, mayor es la variabili-

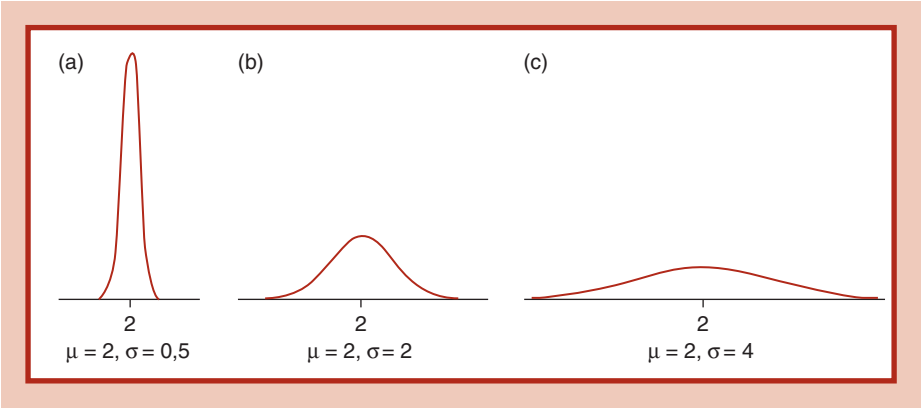


Figura 6.4 Tres funciones de densidad de probabilidad normal.

dad de la variable. En la figura 6.3 se presentan tres funciones distintas de densidad de probabilidad normal. Observe que estas curvas se hacen más planas a medida que σ crece.

Debido a que la función de densidad de probabilidad de una variable aleatoria normal es simétrica respecto a su valor esperado μ, se tiene que es igualmente probable que X se encuentre por encima o por debajo de μ. Esto es,

$$P\{X < \mu\} = P\{X > \mu\} = \frac{1}{2}$$

No todas las curvas de densidad con forma acampanada son normales. Las curvas de densidad normales vienen determinadas por una fórmula específica: la altura de la curva por encima del punto con una abscisa x es

$$\frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2/2\sigma^2}$$

Aunque esta fórmula no se utilizará directamente es interesante observar que incluye dos famosas constantes matemáticas: π (el área del círculo de radio 1) y e (que es la base de los logaritmos naturales). Igualmente, observe que esta fórmula queda completamente especificada si se conoce el valor medio μ y la desviación típica σ.

Una variable aleatoria normal con media 0 y varianza 1 se denomina variable aleatoria *normal estándar*, y su curva de densidad se conoce como *curva de densidad normal estándar*. La figura 6.5 muestra la función de densidad normal estándar. En este texto se usará (y se reservará) la letra Z para representar la variable aleatoria normal estándar.

En la sección 6.5, se verá cómo se pueden determinar las probabilidades asociadas a una variable aleatoria normal arbitraria, relacionándolas con las probabilidades correspondientes a una variable aleatoria normal estándar. Para hacer esto, se utilizará la siguiente regla de aproximación de las probabilidades normales.

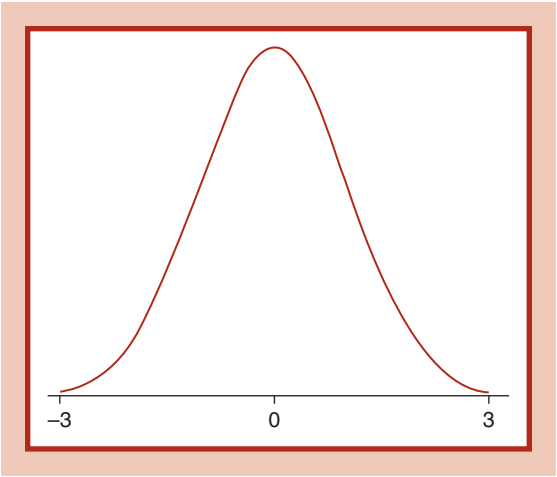


Figura 6.5 Curva normal estándar.

Regla de aproximación

Una variable aleatoria normal con media μ y desviación típica σ estará:

Entre $\mu - \sigma$ y $\mu + \sigma$ con una probabilidad aproximada de 0,68

Entre $\mu - 2\sigma$ y $\mu + 2\sigma$ con una probabilidad aproximada de 0,95

Entre $\mu - 3\sigma$ y $\mu + 3\sigma$ con una probabilidad aproximada de 0,997

Esta regla de aproximación viene reflejada en la figura 6.6. A menudo, esta regla permite que nos hagamos una idea sobre si un determinado conjunto de datos se aproxima o no a una normal.

Ejemplo 6.1 Las calificaciones obtenidas en un test de aptitud oral se distribuyen normalmente con media igual a 504. Si la desviación típica de las calificaciones es 84, se verifica que, aproximadamente, el 68% de todas las calificaciones se encuentran entre $504 - 84$ y $504 + 84$. Esto es, aproximadamente el 68% de las calificaciones se encuentran entre 420 y 588. De igual forma, aproximadamente el 95% de las calificaciones se encuentran entre $504 - 168 = 336$ y entre $504 + 168 = 672$; y aproximadamente un 99,7% se encuentran entre 252 y 756. ■

La citada regla de aproximación constituye la base teórica de la regla empírica que se presentará en la sección 3.6. La relación entre ambas reglas se clarificará en el capítulo 8, cuando se vea cómo se pueden utilizar la media muestral y la desviación típica muestral para estimar μ y σ.

Si se tiene en cuenta la simetría de la curva normal respecto del valor μ, se pueden obtener otras muchas conclusiones a partir de la aproximación normal. Por ejemplo, dado que el área entre μ y $\mu + \sigma$ es igual a la comprendida entre $\mu - \sigma$ y μ, se desprende de esta regla que una variable aleatoria normal se encontrará entre μ y $\mu + \sigma$ con una probabilidad aproximada de $0{,}68/2 = 0{,}34$.

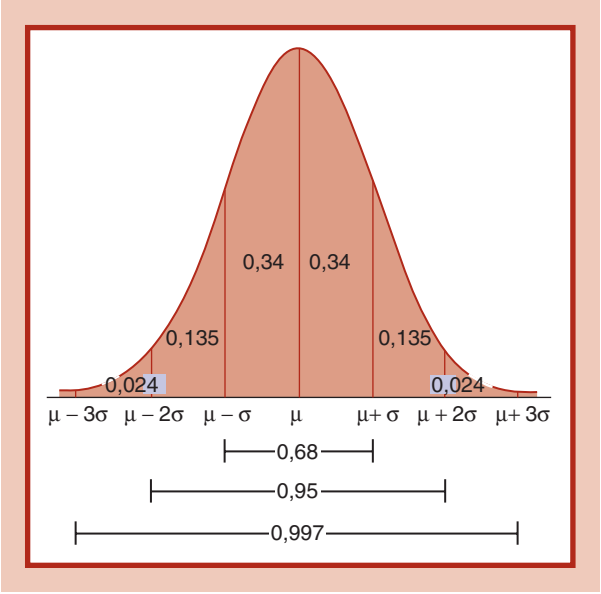


Figura 6.6 Áreas aproximadas bajo la curva normal.

Problemas

1. Las presiones sanguíneas sistólicas de los adultos, en las unidades apropiadas, se distribuyen normalmente con media 128,4 y desviación típica 19,6.

   (a) Obtenga un intervalo que contenga las presiones sanguíneas de aproximadamente el 68% de la población.

   (b) Calcule un intervalo contenga las presiones sanguíneas de aproximadamente el 95% de la población.

   (c) Obtenga un intervalo contenga las presiones sanguíneas de aproximadamente el 99,7% de la población.

2. Las alturas de los varones de una población se distribuyen normalmente con media 69 pulgadas y desviación típica de 6,5 pulgadas. Aproxime la proporción de la población con una altura inferior a 82 pulgadas.

Los problemas del 3 al 16 son de elección múltiple. Especifique la respuesta que, en su opinión, se aproxima más a la respuesta verdadera. Recuerde que Z representa una variable normal estándar. Dibuje un gráfico que en cada caso justifique su contestación.

3. $P\{-2 < Z < 2\}$ es aproximadamente:

   (a) 0,68 (b) 0,95 (c) 0,975 (d) 0,50

4. $P\{Z > -1\}$ es aproximadamente:

   (a) 0,50 (b) 0,95 (c) 0,84 (d) 0,16

5. $P\{Z > 1\}$ es aproximadamente:

   (a) 0,50 (b) 0,95 (c) 0,84 (d) 0,16

6. $P\{Z > 3\}$ es aproximadamente:

   (a) 0,30 (b) 0,05 (c) 0 (d) 0,99

7. $P\{Z < 2\}$ es aproximadamente:

   (a) 0,95 (b) 0,05 (c) 0,975 (d) 0,025

En los problemas del 8 al 11, X es una variable aleatoria normal con valor esperado 15 y desviación estándar 4.

8. La probabilidad de que X esté entre 11 y 19 es aproximadamente:

   (a) 0,50 (b) 0,95 (c) 0,68 (d) 0,34

9. La probabilidad de que X sea menor que 23 es aproximadamente:

   (a) 0,975 (b) 0,95 (c) 0,68 (d) 0,05

10. La probabilidad de que X sea menor que 11 es aproximadamente:

    (a) 0,34 (b) 0,05 (c) 0,16 (d) 0,50

11. La probabilidad de que X sea mayor que 27 es aproximadamente:

    (a) 0,05 (b) 0 (c) 0,01 (d) 0,32

12. La variable X es una variable aleatoria normal con desviación típica 3. Si la probabilidad de que X esté entre 7 y 19 es 0,95, el valor esperado de X es aproximadamente:

    (a) 16 (b) 15 (c) 14 (d) 13

13. La variable X es una variable aleatoria normal con desviación típica 3. Si la probabilidad de que X sea menor que 16 es 0,84, el valor esperado de X será aproximadamente:

    (a) 16 (b) 15 (c) 14 (d) 13

14. La variable X es una variable aleatoria normal con desviación típica 3. Si la probabilidad de que X sea mayor que 16 es 0,975, el valor esperado de X será aproximadamente:

    (a) 20 (b) 22 (c) 23 (d) 25

15. La variable X es una variable aleatoria normal con valor esperado 100. Si la probabilidad de que X sea menor que 90 es 0,84, la desviación típica de X será aproximadamente:

    (a) 5 (b) 10 (c) 15 (d) 20

16. La variable X es una variable aleatoria normal con un valor esperado 100. Si la probabilidad de que X sea mayor que 130 es 0,025, la desviación típica de X será aproximadamente:

    (a) 5 (b) 10 (c) 15 (d) 20

17. La variable X es una variable aleatoria normal con valor esperado 100 y desviación típica 2, e Y es una normal con valor esperado 105 y desviación típica 10. Calcule cuál de las dos variables, X o Y, tiene mayor probabilidad de que:

    (a) Sobrepase 104. (b) Sobrepase 96. (c) Sobrepase 100.

18. La variable X es una variable aleatoria normal con valor esperado 100 y desviación típica 2, e Y es una normal con valor esperado 105 y desviación típica 10. Calcule cuál de las dos variables, X o Y, tiene mayor probabilidad de que:

    (a) Sobrepase 105. (b) Sea menor que 95.

19. Las calificaciones de un determinado test de aptitud en el trabajo son normales con valor esperado 400 y desviación típica 100. Si una compañía considera solamente aquellas solicitudes de trabajo cuyas calificaciones en el test se encuentren en el 5% más alto, determine si la compañía considerará una solicitud cuya puntuación en el test haya sido:

    (a) 400 (b) 450 (c) 500 (d) 600

6.4 Probabilidades asociadas a la variable aleatoria normal estándar

Sea Z una variable aleatoria normal estándar. Esto es, Z es una variable aleatoria normal con media 0 y desviación típica 1. La probabilidad de que Z esté entre dos valores a y b es igual al área bajo la curva normal estándar entre a y b. Se han computado las áreas bajo esta curva y se han publicado las tablas que nos permiten encontrar las probabilidades de intervalos. Una de las tablas citadas es la tabla 6.1.

Para cada valor no negativo de x, la tabla 6.1 especifica la probabilidad de que Z sea menor que x. Por ejemplo, supongamos que se desea determinar $P\{Z < 1{,}22\}$. Esto se puede hacer buscando en la entrada de la tabla por filas hasta encontrar la fila con valor 1,2, y luego buscando en la entrada de la tabla por columnas hasta encontrar la columna con entrada 0,02. El valor correspondiente a la fila con entrada 1,2 y a la columna con valor 0,02 muestra la probabilidad pedida. Puesto que este valor es 0,8888, se ve que

$$P\{Z < 1{,}22\} = 0{,}8888$$

Tabla 6.1 Probabilidades de la normal estándar.

x	**0,00**	**0,01**	**0,02**	**0,03**	**0,04**	**0,05**	**0,06**	**0,07**	**0,08**	**0,09**
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998

El valor representado en la tabla es $P\{Z < x\}$.

Una parte de la tabla 6.1 que ilustra lo anteriormente dicho se muestra a continuación:

			↓			
x		**0,00**	**0,01**	**0,02**	**0,03**	$\cdots$ **0,09**
	0,0	0,5000	0,5040			
	.					
	.					
	.					
	1,1	0,8413				
→	**1,2**	0,8849	0,8869	0,8888		
	1,3	0,9032				

Se puede utilizar la tabla 6.1 para determinar la probabilidad de que Z sea mayor que x. Por ejemplo, supongamos que se desea determinar la probabilidad de que Z sea mayor que 2. Para hacerlo observe que Z debe ser necesariamente o menor o igual que 2, o bien mayor que 2; por consiguiente

$$P\{Z \leq 2\} + P\{Z > 2\} = 1$$

o, lo que es igual,

$$P\{Z > 2\} = 1 - P\{Z \leq 2\}$$

$$= 1 - 0{,}9772$$

$$= 0{,}0228$$

En otras palabras, la probabilidad de que Z sea mayor que x se puede obtener si se resta de 1 la probabilidad de que Z sea menor que x. Es decir, para cada x,

$$P\{Z > x\} = 1 - P\{Z \leq x\}$$

Ejemplo 6.2 Encuentre:

(a) $P\{Z < 1{,}5\}$

(b) $P\{Z \geq 0{,}8\}$

Solución

(a) De la tabla 6.1, se tiene que

$$P\{Z < 1{,}5\} = 0{,}9332$$

(b) De la tabla 6.1, $P\{Z < 0{,}8\} = 0{,}7881$ y, por tanto,

$$P\{Z \geq 0{,}8\} = 1 - 0{,}7881 = 0{,}2119 \quad \blacksquare$$

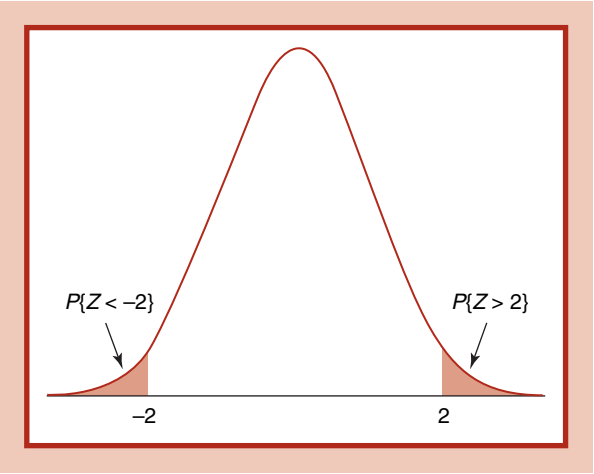


Figura 6.7 $P\{Z < -2\} = P\{Z > 2\}$.

Aunque la tabla 6.1 muestra $P\{Z < x\}$ únicamente para valores no negativos de x, también se puede utilizar cuando x es negativo. Las probabilidades para valores negativos de x se obtienen a partir de la tabla, utilizando la simetría de la curva normal respecto de cero. Por ejemplo, supongamos que se pretende calcular la probabilidad de que Z sea menor que -2. Por la simetría (véase la figura 6.7), la probabilidad citada es igual a la probabilidad de que Z sea mayor que 2; en consecuencia:

$$\begin{aligned} P\{Z< -2\} &= P\{Z>2\} \\ &= 1 - P\{Z<2\} \\ &= 1 - 0{,}9772 = 0{,}0028 \end{aligned}$$

En general, para cualquier valor de x,

$$P\{Z < -x\} = P\{Z > x\} = 1 - P\{Z < x\}$$

Se puede determinar la probabilidad de que Z esté comprendida entre a y b, con $a < b$, si se determina primero la probabilidad de que Z sea menor que b y después se resta de este valor la probabilidad de que Z sea menor que a (véase la figura 6.8).

Ejemplo 6.3 Encuentre:

(a) $P\{1 < Z < 2\}$

(b) $P\{-1{,}5 < Z < 2{,}5\}$

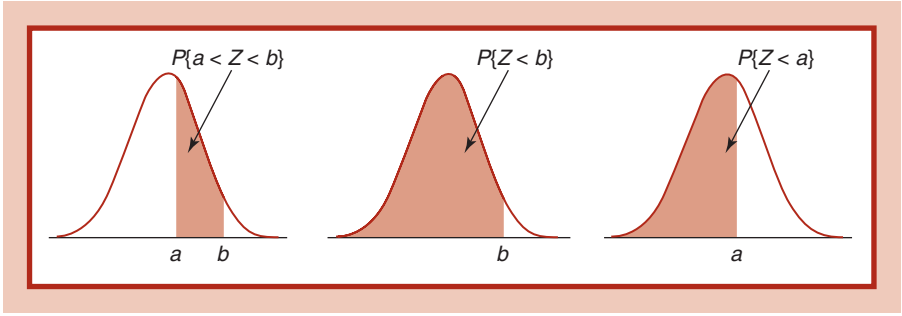


Figura 6.8 $P\{a < Z < b\} = P\{Z < b\} - P\{Z < a\}$.

Solución

(a) $P\{1 < Z < 2\} = P\{Z < 2\} - P\{Z < 1\}$

$= 0{,}9772 - 0{,}8413$

$= 0{,}1359$

(b) $P\{-1{,}5 < Z < 2{,}5\} = P\{Z < 2{,}5\} - P\{Z < -1{,}5\}$

$= P\{Z < 2{,}5\} - P\{Z > 1{,}5\}$

$= 0{,}9938 - (1 - 0{,}9332)$

$= 0{,}9270$

Sea a un valor positivo e intentemos calcular $P\{|Z| > a\}$, la probabilidad de que una normal estándar sea, en valor absoluto, mayor que a. Puesto que $|Z|$ es mayor que a si $Z > a$ o bien si $Z < -a$, se ve que:

$$P\{|Z| > a\} = P\{Z > a\} + P\{Z < -a\}$$

$$= 2P\{Z > a\}$$

donde en la última igualdad se utiliza la simetría de la curva de densidad normal estándar (figura 6.9). ■

Ejemplo 6.4 Encuentre $P\{|Z| > 1{,}8\}$.

Solución

$$P\{|Z| > 1{,}8\} = 2P\{Z > 1{,}8\}$$

$$= 2(1 - 0{,}9641)$$

$$= 0{,}0718 \quad ■$$

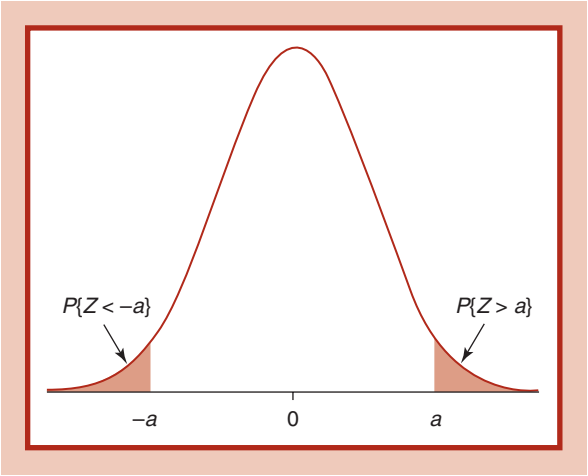


Figura 6.9 $P\{Z > a\} = P\{Z < -a\}$.

Otro resultado que se puede obtener fácilmente es que para cualquier valor positivo a

$$P\{-a < Z < a\} = 2P\{Z < a\} - 1$$

La comprobación de este resultado se deja como ejercicio.

La tabla 6.1 se incluye, también, en el Apéndice D como tabla D.1. Adicionalmente, se puede utilizar el Programa 6-1 para obtener las probabilidades normales. Basta con introducir el valor de x y el programa obtiene como salida $P\{Z < x\}$.

Ejemplo 6.5 Determine $P\{Z > 0{,}84\}$.

Solución Se puede utilizar la tabla 6.1 o bien el Programa 6-1, que computa la probabilidad de que una variable aleatoria normal estándar sea menor que x. Con el Programa 6-1, se puede obtener que, si el valor de x es 0,84, la probabilidad anterior es 0,7995459.

Así pues, la probabilidad pedida en el enunciado es $1 - 0{,}80 = 0{,}20$. Es decir, la probabilidad de que una normal estándar sobrepase el valor 0,84 es aproximadamente igual al 20%. ■

Problemas

1. Para una variable aleatoria normal estándar Z, encuentre:

   (a) $P\{Z < 2{,}2\}$

   (b) $P\{Z > 1{,}1\}$

   (c) $P\{0 < Z < 2\}$

   (d) $P\{-0{,}9 < Z < 1{,}2\}$

(e) $P\{Z > -1{,}96\}$

(f) $P\{Z < -0{,}72\}$

(g) $P\{|Z| < 1{,}64\}$

(h) $P\{|Z| > 1{,}20\}$

(i) $P\{-2{,}2 < Z < 1{,}2\}$

2. Demuestre que $-Z$ también es una variable aleatoria normal estándar. *Sugerencia*: Basta con demostrar que, para todo x,

$$P\{-Z < x\} = P\{Z < x\}$$

3. Encuentre el valor del signo de interrogación:

$$P\{-3 < Z < -2\} = P\{2 < Z < ?\}$$

Utilice un gráfico para verificar que la respuesta es correcta.

4. Utilice la gráfica de la curva normal estándar para comprobar que

$$P\{Z > -2\} = P\{Z < 2\}$$

5. Razone con la ayuda de gráficos o ecuaciones, que para cualquier valor positivo a,

$$P\{-a < Z < a\} = 2P\{Z < a\} - 1$$

6. Encuentre:

(a) $P\{-1 < Z < 1\}$

(b) $P\{|Z| < 1{,}4\}$

7. Encuentre, con dos cifras decimales, el valor de x para el que:

(a) $P\{Z > x\} = 0{,}05$

(b) $P\{Z > x\} = 0{,}025$

(c) $P\{Z > x\} = 0{,}005$

(d) $P\{Z < x\} = 0{,}50$

(e) $P\{Z < x\} = 0{,}66$

(f) $P\{|Z| < x\} = 0{,}99$

(g) $P\{|Z| < x\} = 0{,}75$

(h) $P\{|Z| > x\} = 0{,}90$

(i) $P\{|Z| > x\} = 0{,}50$

6.5 Búsqueda de las probabilidades de la normal: conversión a la normal estándar

Sea X una variable aleatoria normal con media μ y desviación típica σ. Se pueden determinar las probabilidades relativas a X si se utiliza que la variable Z definida por

$$Z = \frac{X - \mu}{\sigma}$$

sigue una distribución normal estándar. Es decir, si se *estandariza* una variable aleatoria normal, restándola su media y dividiéndola por su desviación típica, la variable resultante se convierte en una distribución normal estándar.

El valor de la variable estandarizada nos indica cuánto difiere la variable original de su media en unidades de desviación típica. Por ejemplo, si la variable estandarizada Z toma el valor 2, esto significa que

$$Z = \frac{X - \mu}{\sigma} = 2$$

lo que equivale a

$$X - \mu = 2\sigma$$

Esto es, X es dos desviaciones típicas mayor que su media.

Se puede calcular cualquier probabilidad referida a X si se reescribe equivalentemente en términos de $Z = (X - \mu)/\sigma$ y, después, se utiliza la tabla 6.1 o el Programa 6-1. Por ejemplo, supongamos que se quiere calcular $P\{X < a\}$. Puesto que $X < a$ es equivalente a

$$\frac{X - \mu}{\sigma} < \frac{a - \mu}{\sigma}$$

se ve que

$$P\{X < a\} = P\left\{\frac{X - \mu}{\sigma} < \frac{a - \mu}{\sigma}\right\}$$

$$= P\left\{Z < \frac{a - \mu}{\sigma}\right\}$$

donde Z representa una variable aleatoria normal estándar.

Ejemplo 6.6 Las puntuaciones obtenidas en un test de inteligencia por los alumnos de sexto curso de primaria se distribuyen normalmente con media 100 y desviación típica 14,2.

(a) ¿Cuál es la probabilidad de que un alumno de sexto curso de primaria elegido aleatoriamente obtenga una puntuación mayor que 130 en el citado test?

(b) ¿Cuál es la probabilidad de que el alumno seleccionado obtenga una puntuación comprendida entre 90 y 115?

Solución Denote por X la puntuación obtenida en el test por el alumno elegido aleatoriamente. Se calcularán las probabilidades relativas a X si se tiene en cuenta que la variable aleatoria estandarizada

$$Z = \frac{X - 100}{14{,}2}$$

sigue una distribución normal estándar.

$$\text{(a) } P\{X > 130\} = P\left\{\frac{X - 100}{14{,}2} > \frac{130 - 100}{14{,}2}\right\}$$

$$= P\{Z > 2{,}1127\}$$

$$= 0{,}017$$

(b) La desigualdad $90 < X < 115$ equivale a

$$\frac{90 - 100}{14{,}2} < \frac{X - 100}{14{,}2} < \frac{115 - 100}{14{,}2}$$

es decir,

$$-0{,}7042 < Z < 1{,}056$$

Por consiguiente,

$$P\{90 < X < 115\} = P\{-0{,}7042 < Z < 1{,}056\}$$

$$= P\{Z < 1{,}056\} - P\{Z < -0{,}7042\}$$

$$= 0{,}854 - 0{,}242$$

$$= 0{,}612 \quad \blacksquare$$

Ejemplo 6.7 Si X es una normal con media μ y desviación típica σ, encuentre:

(a) $P\{|X - \mu| > \sigma\}$

(b) $P\{|X - \mu| > 2\sigma\}$

(c) $P\{|X - \mu| > 3\sigma\}$

Solución Si se tiene en cuenta que la desigualdad $|X - \mu| > a\sigma$ es equivalente, en términos de la variable estandarizada $Z = (X - \mu)/\sigma$, a $|Z| > a$ se obtienen los siguientes resultados:

$$\begin{aligned}(a)\ P\{|X-\mu|>\sigma\} &= P\{|Z|>1\}\\ &= 2P\{Z>1\}\\ &= 2(1-0{,}8413)\\ &= 0{,}3174\end{aligned}$$

$$\begin{aligned}(b)\ P\{|X-\mu|>2\sigma\} &= P\{|Z|>2\}\\ &= 2P\{Z>2\}\\ &= 0{,}0456\end{aligned}$$

$$\begin{aligned}(c)\ P\{|X-\mu|>3\sigma\} &= P\{|Z|>3\}\\ &= 2P\{Z>3\}\\ &= 0{,}0026\end{aligned}$$

Así pues, se ve que la probabilidad de que una variable aleatoria normal difiera de su media en más de una desviación típica es (con dos cifras decimales) igual a 0,32; o equivalentemente, la probabilidad complementaria de que difiera de su media en menos de una desviación típica es 0,68. Similarmente, las partes (b) y (c) implican, respectivamente, que la probabilidad de que la variable aleatoria difiera de su media en menos de dos veces su desviación típica es 0,95 y la de que difiera en menos de 3 veces su desviación típica es 0,997. En consecuencia, se ha comprobado la regla de aproximación presentada en la sección 6.3. ■

6.6 Propiedad aditiva de las variables aleatorias normales

El hecho de que $Z = (X - \mu)/\sigma$ sea una variable aleatoria normal estándar se desprende de la propiedad de que si a una variable aleatoria normal se le suma una constante, o se multiplica por ésta, la variable aleatoria resultante continúa siendo normal. Como consecuencia, si X es una normal con media μ y desviación típica σ, la variable $Z = (X - \mu)/\sigma$ será también normal. Resulta sencillo comprobar que Z tiene media 0 y varianza 1.

Un hecho importante que afecta a las variables aleatorias normales es que la suma de variables aleatorias normales e independientes es igualmente una variable aleatoria normal. Esto es, si las variables aleatorias X e Y son normales e independientes con parámetros respectivos μ_x, σ_x y μ_y, σ_y, $X + Y$ será también normal. Su valor medio es

$$E[X + Y] = E[X] + E[Y] = \mu_x + \mu_y$$

y su varianza es

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) = \sigma_x^2 + \sigma_y^2$$

Es decir, se ha obtenido el siguiente resultado.

Supongamos que X e Y son variables aleatorias normales e independientes con medias μ_x y μ_y y con desviaciones típicas σ_x y σ_y, respectivamente. En estas condiciones, $X + Y$ es normal con media

$$E[X + Y] = \mu_x + \mu_y$$

y con desviación típica

$$\text{SD}(X + Y) = \sqrt{\sigma_x^2 + \sigma_y^2}$$

Ejemplo 6.8 Supongamos que el tiempo que dura una bombilla encendida es una variable aleatoria normal con media 400 horas y desviación típica 40 horas. Si un individuo compra dos bombillas, una de las cuales sirve de repuesto para reemplazar a la otra cuando se funda, ¿cuál es la probabilidad de que con ambas se disponga de más de 750 horas de luz?

Solución Se precisa calcular la probabilidad de que $X + Y > 750$, y siendo X la duración de la primera bombilla e Y la de la segunda. Las variables X e Y son normales con media 400 y desviación típica 40. Adicionalmente, se supondrá que son independientes; por consiguiente, $X + Y$ también será normal con media 800 y desviación típica $\sqrt{40^2 + 40^2} = \sqrt{3200}$. Así pues, $Z = (X + Y - 800)/\sqrt{3200}$ sigue una distribución normal estándar. De esto se desprende que

$$\begin{aligned} P\{X + Y > 750\} &= P\left\{\frac{X + Y - 800}{\sqrt{3200}} > \frac{750 - 800}{\sqrt{3200}}\right\} \\ &= P\{Z > -0{,}884\} \\ &= P\{Z < 0{,}884\} \\ &= 0{,}81 \end{aligned}$$

Por consiguiente, la probabilidad de que se superen 750 horas de luz con ambas bombillas será del 81%. ■

Ejemplo 6.9 Los datos del Departamento de Agricultura de Estados Unidos indican que el consumo de manzanas de una mujer elegida aleatoriamente se distribuye según una normal de media de 19,9 libras y una desviación típica de 3,2 libras, mientras que el consumo de manzanas de un hombre elegido aleatoriamente sigue una normal con media 20,7 libras

y desviación típica 3,4 libras. Supongamos que se elige aleatoriamente a un hombre y una mujer. ¿Cuál es la probabilidad de que el consumo de manzanas de la mujer sea mayor que el del hombre?

Solución Denotemos como X e Y, respectivamente, los consumos de manzanas de la mujer y el hombre seleccionados. Se debe calcular $P\{X > Y\}$ o, equivalentemente, $P\{X - Y > 0\}$. Ahora bien, X es una variable aleatoria normal con media 19,9 y desviación típica 3,2. Adicionalmente, $-Y$ es una variable aleatoria normal (puesto que es igual a la variable aleatoria normal Y multiplicada por la constante -1) con media $-20,7$ y desviación típica $|-1|(3,4) = 3,4$. Por consiguiente, su suma $X + (-Y)$ $(= X - Y)$ sigue una normal con media

$$E[X - Y] = 19,9 + (-20,7) = -0,8$$

y desviación típica

$$\mathrm{SD}(X - Y) = \sqrt{(3,2)^2 + (3,4)^2} = 4,669$$

Así pues, si $W = X - Y$, se tiene que:

$$P\{W > 0\} = P\left\{\frac{W + 0,8}{4,669} > \frac{0,8}{4,669}\right\}$$

$$= P\{Z > 0,17\}$$

$$= 1 - 0,5675 = 0,4325$$

Esto es, la probabilidad de que la mujer elegida aleatoriamente efectúe un consumo de manzanas superior al del hombre elegido aleatoriamente es 0,4325. ■

Problemas

1. Explique detalladamente por qué la desigualdad

$$x > a$$

es equivalente a la desigualdad

$$\frac{x - \mu}{\sigma} > \frac{a - \mu}{\sigma}$$

¿Qué se debe tener en cuenta sobre los posibles valores de σ? (*Sugerencia*: Si σ fuera negativo, ¿serían equivalentes las dos desigualdades?)

2. Si X es normal con media 10 y desviación típica 3, encuentre:

(a) $P\{X > 12\}$

(b) $P\{X < 13\}$

(c) $P\{8 < X < 11\}$

(d) $P\{X > 7\}$

(e) $P\{|X - 10| > 5\}$

(f) $P\{X > 10\}$

(g) $P\{X > 20\}$

3. El tiempo que funciona un secador de pelo antes de que se estropee se distribuye de acuerdo con una normal con media 40 meses y desviación típica 8 meses. El productor estudia la posibilidad de ofrecer a los compradores del secador de pelo una garantía de funcionamiento de 3 años. ¿Cuál será la proporción de secadores que cumplirán la garantía citada?

4. Las puntuaciones obtenidas en un determinado test escolar se distribuyen según una normal con media 520 y desviación típica 94.

   (a) Si la puntuación obtenida por un alumno fue 700, ¿en cuántas veces la desviación típica esta puntuación supera a la puntuación media?

   (b) ¿Cuál es el porcentaje de alumnos que obtienen una puntuación superior a la indicada anteriormente?

5. El número de frascos de champú vendidos mensualmente en una determinada droguería es una variable aleatoria normal con media 212 y desviación típica 40. Encuentre la probabilidad de que las ventas de champú en el próximo mes sean:

   (a) más de 200

   (b) menos de 250

   (c) más de 200 y menos de 250

6. La duración de los neumáticos de un determinado automóvil se distribuye normalmente con media 35 000 millas y desviación típica 5000 millas.

   (a) ¿Qué proporción de neumáticos tendrán una duración comprendida entre 30 000 y 40 000 millas?

   (b) ¿Qué proporción de neumáticos tendrán una duración superior a 40 000 millas?

   (c) ¿Qué proporción de neumáticos aguantarán más de 50 000 millas?

7. Suponga que una persona compra uno de los neumáticos citados en el problema 6. Si el neumático continúa en condiciones aceptables tras circular 40 000 millas, ¿cuál es la probabilidad condicionada de que continúe en condiciones aceptables tras circular otras 10 000 millas adicionales?

8. El pulso del corazón de los jóvenes adultos se distribuye según una normal con media 72 pulsaciones por minuto y desviación típica 9,5 pulsaciones por minuto. Dado que la regulación militar impone no aceptar a reclutas que tengan un número de pulsaciones superior a 95 por minuto, ¿qué porcentaje de la población de jóvenes adultos no cumple el estándar citado?

9. El tiempo que se necesita para rellenar los impresos de solicitud de cierto crédito sigue una distribución normal con media 90 minutos y desviación típica 15 minutos.

Calcule la probabilidad de que un solicitante del crédito tarde en rellenar los impresos:

(a) menos de 75 minutos

(b) más de 100 minutos

(c) entre 90 y 120 minutos

10. Está indicado en las etiquetas de las cajas que los tornillos fabricados por un determinado productor tienen un diámetro comprendido entre 1,09 y 1,11 pulgadas. Si del proceso de producción resulta que el diámetro de los tornillos es una variable aleatoria normal con media 1,10 pulgadas y desviación típica 0,005 pulgadas, ¿qué porcentaje de tornillos no cumple la especificación de las etiquetas?

11. La presión de activación de una válvula producida por una determinada compañía es una variable aleatoria normal con valor esperado 26 libras por pulgada al cuadrado y desviación típica 4 libras por pulgada al cuadrado. ¿Qué porcentaje de las válvulas producidas por la compañía citada tiene una presión de activación comprendida entre 20 y 32 libras por pulgada cuadrada?

12. Una persona planea desguazar su coche viejo después de que recorra otras 20 000 millas. La batería del coche acaba de fallar, y la persona debe decidir qué tipo de batería, entre dos posibles que cuestan lo mismo, debe adquirir. Tras realizar ciertas averiguaciones descubre que la primera batería se distribuye normalmente con una vida media de 24 000 millas y con una desviación típica de 6000 millas, mientras que la segunda batería es igualmente normal con media 22 000 millas y desviación típica 2000 millas.

    (a) Si lo único que preocupa a dicha persona es que la batería que compre dure al menos 20 000 millas, ¿cuál debería adquirir?

    (b) ¿Qué ocurriría si la citada persona quisiera que la batería durase 21 000 millas?

13. El tiempo de vida de una televisión es una variable aleatoria normal con media 8,2 años y desviación típica 1,4 años. Calcule el porcentaje de televisores que duran:

    (a) más de 10 años

    (b) menos de 5 años

    (c) entre 5 y 10 años

14. La cantidad de lluvia caída anualmente en Cincinnati, Ohio, se distribuye normalmente con media 40,14 pulgadas y desviación típica 8,7 pulgadas.

    (a) ¿Cuál es la probabilidad de que este año caigan más de 42 pulgadas de lluvia?

    (b) ¿Cuál es la probabilidad de que la lluvia total que caiga en los 2 años próximos sobrepase las 84 pulgadas?

    (c) ¿Cuál es la probabilidad de que la lluvia total que caiga en los 3 años próximos sobrepase las 126 pulgadas?

    (d) En los apartados (b) y (c), ¿qué hipótesis de independencia se asume?

15. La altura de las mujeres adultas de Estados Unidos se distribuye según una normal con media 64,5 pulgadas y desviación típica 2,4 pulgadas. Calcule la probabilidad de que una mujer elegida aleatoriamente tenga una altura de:

    (a) menos de 64 pulgadas

    (b) menos de 70 pulgadas

    (c) entre 63 y 74 pulgadas

    (d) Alicia tiene una altura de 72 pulgadas. ¿Qué porcentaje de mujeres tienen menos altura que Alicia?

    (e) Encuentre la probabilidad de que el promedio de las alturas de dos mujeres elegidas aleatoriamente esté por encima de 67,5 pulgadas.

16. Los pesos de los libros de texto de introducción a la Química son una variable aleatoria con media 3,5 libras y desviación típica 2,2 libras, mientras que los pesos de los libros de texto de introducción a la Economía siguen una normal con media 4,6 libras y desviación típica 1,3 libras. Si Alicia pretende matricularse en cursos de introducción a la Química y a la Economía, calcule la probabilidad de que:

    (a) El peso total de sus dos libros de texto sobrepase las 9 libras.

    (b) Su libro de Economía pese más que su libro de Química.

    (c) ¿Qué hipótesis se debe hacer?

6.7 Percentiles de las variables aleatorias normales

Para cualquier valor α comprendido entre 0 y 1, definamos z_α como aquel valor para el que

$$P\{Z > z_\alpha\} = \alpha$$

Dicho con palabras, la probabilidad de que una variable aleatoria normal estándar sea mayor que z_α es igual a α (véase la figura 6.10).

Se puede determinar el valor de z_α mediante la tabla 6.1. Por ejemplo, supongamos que se pretende encontrar $z_{0,025}$. Puesto que

$$P\{Z < z_{0,025}\} = 1 - P\{Z > z_{0,025}\} = 0{,}975$$

se debe buscar en el cuerpo de la tabla 6.1 el valor 0,975 para, después, buscar el x que corresponde a dicho valor. Puesto que el valor 0,975 corresponde a la fila con la entrada 1,9 y a la columna con la entrada 0,06, se ve que

$$z_{0,025} = 1{,}96$$

Esto es, un 2,5% de las veces que se observe una normal estándar se obtendrán valores por encima de 1,96.

Puesto que el 97,5% de las veces que se observe una normal estándar se obtendrán valores inferiores a 1,96, se dice que 1,96 es el percentil de orden 97,5% de la distribución normal estándar. En general, dado que el $100(1 - \alpha)$ por ciento de las veces que se observa una normal estándar el valor observado es inferior a z_α, se dice que z_α es el percentil de orden $100(1 - \alpha)$ por ciento de la distribución normal estándar.

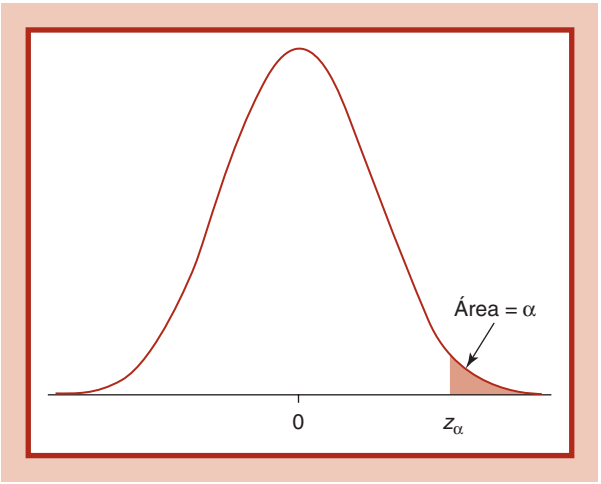


Figura 6.10 $P\{Z > z_\alpha\} = \alpha$.

El valor z_α se denomina *percentil* de orden $100(1 - \alpha)$ por ciento de la distribución normal estándar.

Supongamos ahora que se quiere encontrar $z_{0,05}$. Si se busca en el cuerpo de la tabla 6.1 el valor 0,95, no se puede encontrar este valor exactamente. De hecho, se ve que

$$P\{Z < 1{,}64\} = 0{,}9495$$

y

$$P\{Z < 1{,}65\} = 0{,}9505$$

Por consiguiente, parece que $z_{0,05}$ coincide, más o menos, con el punto medio de 1,64 y 1,65; así pues, lo aproximaremos por 1,645. De hecho, resulta que esta respuesta es la correcta con tres cifras decimales y, por tanto,

$$z_{0,05} = 1{,}645$$

Los valores $z_{0,10}$, $z_{0,05}$, $z_{0,025}$, $z_{0,01}$ y $z_{0,005}$ tienen, como se verá en los siguientes capítulos, una particular importancia en Estadística. Sus valores son los siguientes:

$$z_{0,10} = 1{,}282 \qquad z_{0,025} = 1{,}960 \qquad z_{0,005} = 2{,}576$$
$$z_{0,05} = 1{,}645 \qquad z_{0,01} = 2{,}326$$

Para los restantes valores de α, se puede utilizar la tabla 6.1 para obtener z_α, si se busca la fila y la columna que corresponden al valor más próximo a $1 - \alpha$. También se puede utilizar el Programa 6-2 para obtener z_α.

Ejemplo 6.10 Calcule:

(a) $z_{0,25}$

(b) $z_{0,80}$

Solución

(a) El percentil de orden 75%, $z_{0,25}$, se obtiene a partir de la fila y la columna correspondiente al valor 0,7486 y es aproximadamente igual a 0,67. Así pues, se ve que

$$P\{Z > z_{0,25}\} = 0,25$$

o, equivalentemente,

$$P\{Z < z_{0,25}\} = 0,75$$

La entrada más próxima a 0,75 en la tabla 6.1 es 0,7486, que se corresponde con el valor 0,67. Así pues, se ve que

$$z_{0,25} \approx 0,67$$

Se puede obtener un valor más preciso para $z_{0,25}$ si se utiliza del Programa 6-2. Se obtiene lo siguiente: si a es igual a 0,25, el valor de $z_{0,25}$ es 0,6744897.

(b) Se nos pide encontrar el valor $z_{0,80}$ tal que

$$P\{Z > z_{0,80}\} = 0,80$$

En esta ocasión, el valor $z_{0,80}$ será negativo (¿por qué?); por esta razón, es mejor escribir la ecuación equivalente (véase la figura 6.11)

$$P\{Z < -z_{0,80}\} = 0,80$$

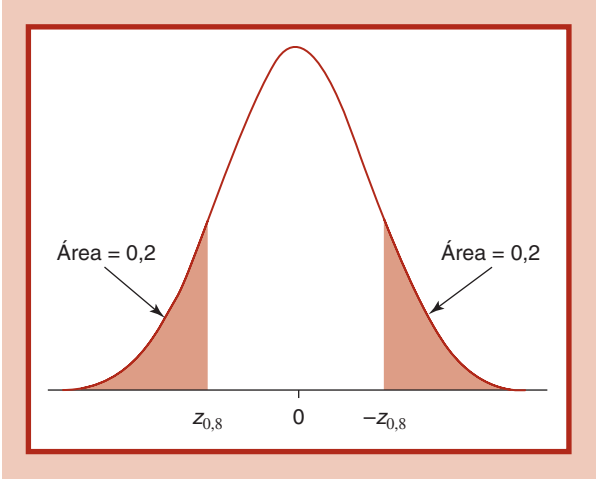


Figura 6.11 $P\{Z < -z_{0,8}\} = 0,80$.

De la tabla 6.1, se ve que

$$-z_{0,80} \approx 0{,}84$$

y, por consiguiente,

$$z_{0,80} \approx -0{,}84 \quad \blacksquare$$

Se pueden obtener los percentiles de cualquier variable aleatoria normal si se convierte en un una normal estándar. Por ejemplo, supongamos que se quiere encontrar el valor de x para el que

$$P\{X < x\} = 0{,}95$$

donde X representa una normal de media 40 y de desviación típica 5. Si se escribe la desigualdad $X < x$ en términos de la variable estandarizada $Z = (X - 40)/5$, se ve que

$$\begin{aligned} 0{,}95 &= P\{X < x\} \\ &= P\left\{\frac{X - 40}{5} < \frac{x - 40}{5}\right\} \\ &= P\left\{Z < \frac{x - 40}{5}\right\} \end{aligned}$$

Ahora bien, $P\{\{Z < z_{0,05}\} = 0{,}95$; por consiguiente, se sigue que

$$\frac{x - 40}{5} = z_{0,05} = 1{,}645$$

de donde, el valor pedido de x es

$$x = 5(1{,}645) + 40 = 48{,}225$$

Ejemplo 6.11 Las calificaciones de un test de inteligencia se distribuyen de acuerdo con una normal de media 100 y desviación típica 14,2. ¿En qué rango de valores se encuentran el 1% más alto de las puntuaciones?

Solución Se debe encontrar el valor de x para el que

$$P\{X > x\} = 0{,}01$$

X es una normal con media 100 y desviación típica 14,2. Ahora bien

$$\begin{aligned} P\{X > x\} &= P\left\{\frac{X - 100}{14{,}2} > \frac{x - 100}{14{,}2}\right\} \\ &= P\left\{Z > \frac{x - 100}{14{,}2}\right\} \end{aligned}$$

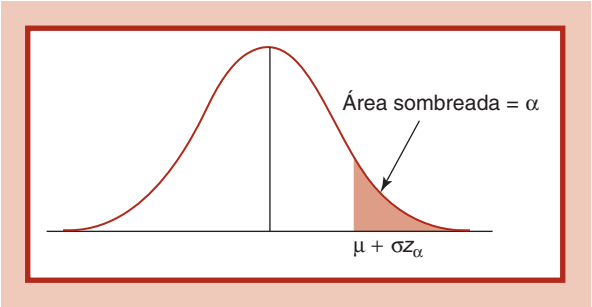


Figura 6.12 $P\{X > \mu + \sigma z_\alpha\} = \alpha$.

Puesto que $P\{Z > z_{0,01}\} = 0,01$, se tiene que la probabilidad anterior referida a X será igual a 0,01 si

$$\frac{x - 100}{14,2} = z_{0,01} = 2,33$$

así pues,

$$x = 14,2(2,33) + 100 = 133,086$$

Esto es, el 1% más alto de las calificaciones se encuentran por encima de 134. ■

La figura 6.12 ilustra el resultado

$$P\{X > \mu + \sigma z_\alpha\} = \alpha$$

cuando X es una variable aleatoria normal con media μ y desviación típica σ.

Problemas

1. Calcule con dos cifras decimales:

   (a) $z_{0,07}$

   (b) $z_{0,12}$

   (c) $z_{0,30}$

   (d) $z_{0,03}$

   (e) $z_{0,65}$

   (f) $z_{0,50}$

   (g) $z_{0,95}$

   (h) $z_{0,008}$

2. Encuentre el valor de x para el que:
   (a) $P\{|Z| > x\} = 0{,}05$
   (b) $P\{|Z| > x\} = 0{,}025$
   (c) $P\{|Z| > x\} = 0{,}005$
3. Si X es una variable aleatoria normal con media 50 y desviación típica 6, calcule el valor aproximado de x para el cual:
   (a) $P\{X > x\} = 0{,}5$
   (b) $P\{X > x\} = 0{,}10$
   (c) $P\{X > x\} = 0{,}025$
   (d) $P\{X < x\} = 0{,}05$
   (e) $P\{X < x\} = 0{,}88$
4. Las puntuaciones de un examen para agentes inmobiliarios se distribuyen normalmente con media 420 y desviación típica 66. Si el tribunal del examen desea dar la calificación de *excelente* al 10% de las puntuaciones más altas, ¿a partir de que puntuación se otorgará la calificación de excelente?
5. Suponga que, en el problema 4, el tribunal pretende que solamente el 25% con puntuaciones más altas superen el examen. ¿Cuál debería ser la puntuación de corte?
6. Los tiempos que tardan los estudiantes de un instituto en correr 1 milla se distribuyen normalmente con media 460 segundos y desviación típica 40 segundos. Se considera que todos aquellos cuyos tiempos están en el 20% más bajo necesitan un entrenamiento adicional. ¿Cuál es el tiempo por debajo del cual se asignarán los entrenamientos adicionales citados?
7. En el problema 6, el 5% más rápido corre 1 milla en menos de x segundos. ¿Cuál es el menor valor de x para el que la sentencia anterior es cierta?
8. Repita el problema 7 reemplazando el *5% más rápido* por el *1% más rápido*.
9. La cantidad de radiación que un individuo puede absorber antes de que le sobrevenga la muerte varía de un individuo a otro. Sin embargo, sobre la población al completo esta cantidad se distribuye normalmente con media 500 roentgens y desviación típica 150 roentgens. ¿Por encima de qué dosis de radiación solamente sobreviviría el 5% de la población?
10. La duración de la transmisión de un coche nuevo se distribuye normalmente con media 70 000 millas y desviación típica 10 000 millas. La compañía productora desea establecer una garantía para dicha transmisión. Si la compañía pretende que solamente el 20% de los coches vendidos puedan revindicar la garantía, ¿cuál debe ser la duración (en millas) establecida en el periodo de garantía?
11. El número de espectadores que asisten a los partidos jugados en casa del equipo de una determinada universidad es una variable aleatoria normal con media 52 000 y desviación típica 4000. ¿Cuáles de las siguientes sentencias son ciertas?

(a) Más de un 80% de los partidos tienen un número de espectadores que sobrepasa los 46 000.

(b) Menos de un 10% de los partidos cuentan con un número de espectadores mayor que 58 000.

12. Las calificaciones de la parte cuantitativa del examen de graduación se distribuyen normalmente con una calificación media 510 y desviación típica 92. Calcule cuál es la calificación que se necesita obtener para estar en el

(a) 10% superior de todas las calificaciones

(b) 5% superior de todas las calificaciones

(c) 1% superior de todas las calificaciones

13. El nivel de glucosa en sangre (por 100 mililitros de sangre) de los diabéticos se distribuye normalmente con media 106 miligramos y desviación típica 8 miligramos. ¿Por debajo de qué valor se debe encontrar el nivel de glucosa de un diabético para que forme parte del 20% de los niveles más bajos?

Términos clave

Variable aleatoria continua: Variable aleatoria que puede tomar cualquier valor contenido en un intervalo.

Función de densidad de probabilidad: Curva asociada a una variable aleatoria continua. La probabilidad de que la variable aleatoria esté comprendida entre dos puntos es igual al área bajo la curva entre dichos puntos.

Variable aleatoria normal: Tipo de variables aleatorias continuas cuyas funciones de densidad de probabilidad son simétricas con formas acampanadas.

Variable aleatoria normal estándar: Variable aleatoria normal con media 0 y varianza 1.

Percentil de orden 100*p* por ciento de una variable aleatoria continua: La probabilidad de que la variable aleatoria sea menor que dicho percentil es *p*.

Resumen

Una variable aleatoria *continua* es aquella que puede tomar cualquier valor comprendido dentro de un intervalo. Sus probabilidades se pueden obtener a partir de su *función de densidad de probabilidad*. Concretamente, la probabilidad de que la variable aleatoria caiga entre los puntos a y b es igual al área que cae por debajo de la función de densidad entre los puntos a y b.

Una variable aleatoria *normal* X tiene una función de densidad de probabilidad que viene determinada por dos parámetros, la media μ y la desviación típica σ de X. La función de densidad tiene una forma acampanada que es simétrica respecto de μ y cuya dispersión crece a medida que σ aumenta.

Una variable aleatoria normal toma valores separados de la media en menos de una vez su desviación típica en aproximadamente un 68% de los casos; toma valores que distan de la media en menos de dos veces su desviación típica en aproximadamente un 95% de los casos; y toma valores separados de la media en menos de tres veces su desviación típica en aproximadamente un 99,7% de los casos.

Una variable aleatoria normal con media 0 y desviación típica 1 se denomina variable aleatoria *normal estándar*. Habitualmente, esta variable se designa con la letra Z. Las probabilidades correspondientes a la variable aleatoria normal estándar se pueden obtener a partir de la tabla 6.1 (reimpresa como tabla D.1 en el Apéndice). Para cualquier valor no negativo x, esta tabla muestra, con dos cifras decimales, la probabilidad de que una variable aleatoria normal estándar sea menor que x. Para valores negativos de x, esta probabilidad se puede obtener si se tiene en cuenta la simetría de la función de densidad normal estándar con respecto a 0. Esto se concreta en la igualdad

$$P\{Z < x\} = P\{Z > -x\}$$

El valor de $P\{Z > -x\} = 1 - P\{Z < -x\}$ puede obtenerse a partir de la tabla 6.1.

El Programa 6-1 también se puede usar para obtener las probabilidades asociadas a las variables aleatorias normales estándar.

Si X es normal con media μ y desviación típica σ, la variable aleatoria Z, definida por

$$Z = \frac{X - \mu}{\sigma}$$

Perspectiva histórica

Bettmann

Karl F. Gauss

La curva normal

La distribución normal fue introducida por el matemático francés Abraham De Moivre en 1733. De Moivre, que usó esta distribución para aproximar probabilidades relacionadas con lanzamientos de monedas, la denominó *curva exponencial con forma acampanada*. Su utilidad, sin embargo, se hizo evidente en 1809, cuando el famoso matemático alemán K. F. Gauss la utilizó para predecir la localización de cuerpos celestes. Como resultado, tras esta fecha se extendió su denominación como *distribución gaussiana.*

Durante la segunda mitad del siglo XIX, la mayor parte de los estadísticos comenzaron a creer que la mayor parte de los conjuntos de datos presentaban histogramas que seguían la forma acampanada gaussiana. Realmente, llegó a aceptarse que era "normal" que cualquier conjunto de datos habitual siguiera esta curva. Como resultado, siguiendo la indicación de Karl Pearson, la gente empezó a denotar la curva gaussiana como simplemente curva *normal*. (Para explicar por qué existen tantos conjuntos de datos que parecen seguir la curva normal, los estudiantes interesados deberán esperar a leer las secciones 7.3 y 12.6.)

Karl Friedrich Gauss (1777-1855), uno de los primeros usuarios de la curva normal, fue uno de los más grandes matemáticos de todos los tiempos. Observe las palabras del conocido historiador de las matemáticas E. T. Bell, tal como las expresó en su libro

Hombres de las Matemáticas. En un capítulo titulado "El Príncipe de los Matemáticos" él escribe:

> *"Arquímedes, Newton y Gauss; estos tres están por sí mismos dentro la clase de los grandes matemáticos, y no es posible para los mortales ordinarios intentar igualarles en mérito. Los tres aportaron ideas relevantes tanto en la matemática pura como en la aplicada. Arquímedes estimó sus matemáticas puras muy por encima de sus aplicaciones; Newton basó la importancia de sus ideas matemáticas en la utilidad científica que tenían; por su parte, Gauss declaró que para él trabajar en el campo puro o en el aplicado era todo lo mismo."*

se distribuye como una normal estándar. Este hecho nos permite calcular las probabilidades de X transformándolas en probabilidades asociadas a Z. Por ejemplo,

$$P\{X < a\} = P\left\{\frac{X - \mu}{\sigma} < \frac{a - \mu}{\sigma}\right\}$$

$$= P\left\{Z < \frac{a - \mu}{\sigma}\right\}$$

Para cualquier valor de a comprendido entre 0 y 1, el valor z_α se define como aquél para el que

$$P\{Z > z_\alpha\} = \alpha$$

Así pues, una normal estándar será menor que z_α con una probabilidad $1 - \alpha$. Es decir, un $100(1 - \alpha)$ por ciento de las veces Z será menor que z_α. El valor z_α se denomina *percentil* de orden $100(1 - \alpha)$ por ciento de la distribución normal estándar.

Se pueden obtener los valores de z_α, para valores determinados de α, bien a partir de la tabla 6.1 o mediante el Programa 6-2. Los percentiles de una variable aleatoria, X, normal arbitraria con media μ y desviación típica σ se pueden obtener usando el hecho de que $Z = (X - \mu)/\sigma$ sigue una distribución normal estándar. Por ejemplo, supongamos que se desea encontrar el valor de x para el que

$$P\{X > x\} = \alpha$$

Esto significa que se pretende encontrar el valor de x para el que

$$\alpha = P\left\{\frac{X - \mu}{\sigma} > \frac{x - \mu}{\sigma}\right\}$$

$$= P\left\{Z > \frac{x - \mu}{\sigma}\right\}$$

Por consiguiente, dado que $P\{Z > z_\alpha\} = \alpha$, se puede concluir que

$$\frac{x - \mu}{\sigma} = z_\alpha$$

o

$$x = \mu + \sigma z_\alpha$$

Problemas de repaso

1. Las alturas de los hombres adultos se distribuyen normalmente con media 69 pulgadas y desviación típica 2,8 pulgadas. Denote por X la altura de un varón adulto elegido aleatoriamente. Calcule:
   (a) $P\{X > 65\}$
   (b) $P\{62 < X < 72\}$
   (c) $P\{|X - 69| > 6\}$
   (d) $P\{63 < X < 75\}$
   (e) $P\{X > 72\}$
   (f) $P\{X < 60\}$
   (g) x si $P\{X > x\} = 0{,}01$
   (h) x si $P\{X < x\} = 0{,}95$
   (i) x si $P\{X < x\} = 0{,}40$
2. Calcule:
   (a) $z_{0,04}$
   (b) $z_{0,22}$
   (c) $P\{Z > 2{,}2\}$
   (d) $P\{Z < 1{,}6\}$
   (e) $z_{0,78}$
3. En experimentos hechos con pilotos de aviación, se encontró que los umbrales de desmayo frente a aceleraciones se distribuyen normalmente con media 4,5 g y desviación típica 0,7 g. Si solamente a los pilotos cuyos umbrales se encuentran en el 25% más alto se les permite ser candidatos a astronauta, ¿cuál es el punto de corte para poder optar a ser astronauta?
4. En el problema 3, encuentre la proporción de pilotos de aviación que tienen sus umbrales de desmayo:
   (a) por encima de 5 g
   (b) por debajo de 4 g
   (c) comprendidos entre 3,7 g y 5,2 g
5. La duración de un tipo de bombilla se distribuye como una normal con media 500 horas y desviación típica 60 horas.
   (a) ¿Cuál es la proporción de bombillas que duran más de 560 horas?
   (b) ¿Cuál es la proporción de bombillas que duran menos de 440 horas?
   (c) Si una bombilla continúa funcionando tras haber estado encendida durante 440 horas, ¿cuál es la probabilidad condicionada de que dure más de 560 horas?

(d) Rellene el número que falta en la siguiente sentencia. El 10% de las bombillas citadas durarán al menos_____ horas.

6. La Sociedad del Cáncer Americana mantiene que un varón de 25 años que fume un paquete de cigarrillos al día reduce su vida 5,5 años, en promedio. Asumiendo que el número de años en que se ve reducida la vida de los fumadores citados sigue una normal con media 5,5 y desviación típica 1,5, calcule la probabilidad de que la disminución de la vida de un fumador:

   (a) Sea menor de 2 años.

   (b) Sea mayor de 8 años.

   (c) Esté comprendida entre 4 y 7 años.

7. Supongamos que los costes anuales de mantenimiento que deben afrontar los propietarios de los apartamentos de un complejo residencial se distribuyen normalmente con media 3000 $ ydesviación típica 600 $. Calcule la probabilidad de que el coste total que debe afrontar un propietario particular en los próximos 2 años:

   (a) Sobrepase 5000 $.

   (b) Sea menor de 7000 $.

   (c) Se encuentre comprendido entre 5000 $ y 7000 $.

   Asuma que los costes correspondientes a dos años diferentes son variables aleatorias independientes.

8. Las velocidades de los coches que circulan por las autopistas de New Jersey se distribuyen según una normal con media 60 millas por hora y desviación típica 5 millas por hora. Si la policía de New Jersey sigue la política de multar solamente al 5% de los conductores que circulan a mayor velocidad, ¿a partir de qué umbral de velocidad se comenzará a multar?

9. Las ventas brutas semanales de un vendedor de coches de segunda mano son normales con media 18 800 $ y desviación típica 9000 $.

   (a) ¿Cuál es la probabilidad de que las ventas de la semana próxima sobrepasen los 20 000 $?

   (b) ¿Cuál es la probabilidad de que las ventas semanales sobrepasen los 20 000 $ en las dos semanas próximas?

   (c) ¿Cuál es la probabilidad de que las ventas totales en las dos semanas siguientes sobrepasen los 40 000 $?

   En los apartados (b) y (c) asuma que las ventas producidas en dos semanas distintas son independientes.

10. El número de millas anuales que acumulan los coches pertenecientes a la flota de una gran compañía de alquiler sigue una normal con media 18 000 millas y desviación típica 1700 millas. Al final del año la compañía vende el 80% de los coches y mantiene el 20% que ha circulado menos. ¿Cree que es verosímil que no se venda un coche que haya circulado 17 400 millas en un año?

11. El análisis de las puntuaciones de los partidos de la liga profesional de fútbol americano ha motivado que algunos investigadores mantengan que un equipo que gane por x puntos en la mitad del encuentro sacará a su oponente un número aleatorio de puntos que se distribuye aproximadamente según una normal con media x y desviación típica 14. Así por ejemplo, la diferencia entre los puntos marcados por un equipo que en la mitad del partido vaya ganando por 5 puntos y los puntos marcados por el equipo contrario es una variable aleatoria normal con media 5 y desviación típica 14. Asumiendo que esta teoría es correcta, determine la probabilidad de que:

    (a) Acabe ganando un equipo que vaya ganando por 7 puntos en la mitad del partido.

    (b) Acabe ganando un equipo que vaya perdiendo por 4 puntos en la mitad del partido.

    (c) Acabe perdiendo un equipo que sobrepase en 14 puntos a su contrario en la mitad del partido.

12. Los datos del Departamento de Agricultura de Estados Unidos correspondientes a 1987 indican que el consumo anual de tomates de una mujer elegida aleatoriamente es una variable aleatoria normal con media 14,0 libras y desviación típica 2,7 libras; mientras que el consumo de tomates de un hombre elegido aleatoriamente sigue una normal con media 14,6 libras y desviación típica 3 libras. Supongamos que se elige aleatoriamente a un hombre y una mujer. Calcule la probabilidad de que, en 1987:

    (a) La mujer haya consumido más de 14,6 libras de tomates.

    (b) El hombre haya consumido menos de 14 libras de tomates.

    (c) La mujer haya consumido más de 15 libras de tomates y el hombre menos de 15 libras.

    (d) La mujer haya consumido mayor cantidad de tomates que el hombre.

13. Suponga en el problema 12 que se elige a una persona con igual probabilidad de que se trate de un hombre o de una mujer. Encuentre la probabilidad de que la persona elegida sea:

    (a) Una mujer que en 1987 haya consumido menos de 14 libras de tomates.

    (b) Un hombre que en 1987 haya consumido más de 14 libras de tomates.

7

Distribuciones de los estadísticos asociados al muestreo

Utiliza la Estadística como el borracho utiliza las farolas: como apoyo en lugar de como iluminación.

Andrew Lang (autor escocés)

Yo podría demostrar la existencia de Dios estadísticamente.

George Gallup, U.S. pollster

Se introduce el concepto de muestreo sobre una distribución poblacional. Se estudian la media y la varianza muestrales, y se obtienen sus esperanzas y varianzas. Se presenta el teorema central del límite y se aplica para demostrar que la distribución de la media muestral es aproximadamente normal.

Se muestrea sobre una población finita, en la que algunos de sus miembros presentan o no una determinada característica de interés. Se demuestra que, cuando el tamaño muestral es grande, el número de miembros de la muestra que presentan la característica sigue aproximadamente una variable aleatoria binomial. Se utiliza el teorema central del límite para demostrar que las probabilidades de esta variable aleatoria se pueden aproximar por las probabilidades de una variable aleatoria normal.

Se presenta la distribución de la varianza muestral cuando la distribución poblacional subyacente es normal.

7.1 Preámbulo

Si en un casino un jugador apuesta 1 $ a un número de la ruleta, puede ganar 35 $ si acierta o perder el dólar apostado si falla. Puesto que la rueda de la ruleta tiene 38 posiciones –numeradas como 0, 00, y cada uno de los enteros desde el 1 hasta el 36– se tiene que la probabilidad de que salga el número apostado es 1/38. En consecuencia, la ganancia esperada de la apuesta es

$$E[\text{ganancia}] = 35\left(\frac{1}{38}\right) - 1\left(\frac{37}{38}\right) = -\frac{2}{38} = -0{,}0526$$

Es decir, la pérdida esperada de cada apuesta es aproximadamente de 5,3 céntimos.

Supongamos que el jugador realiza varias apuestas sucesivas. ¿Cuál es la probabilidad de que acabe ganando al final de todas ellas? Claramente esto depende de cuántas apuestas realice. Si se llevan a cabo 100 apuestas, la probabilidad de que al final se vaya ganando es de 0,4916. Si se han realizado 1000 apuestas, esta probabilidad disminuye hasta 0,39. Tras 100 000 apuestas no sólo casi se tiene la certeza de que se irá perdiendo (la probabilidad de ir ganando es aproximadamente de 0,002), sino que, con una probabilidad del 95%, la pérdida media por apuesta será de 5,26 $\pm$ 1,13 céntimos de dólar (léase, 5,26 más o menos 1,13 céntimos). En otras palabras, en este capítulo se aprenderá que, si el jugador realiza un número suficientemente grande de apuestas, la pérdida esperada por apuesta estará próxima a 5,26 céntimos.

7.2 Introducción

Uno de los puntos clave de la Estadística consiste en extraer conclusiones a partir de un conjunto de datos observados. Por lo general, estos datos proceden de una muestra de individuos de una población, y el objetivo será utilizar esta muestra para sacar conclusiones sobre la población total.

Supongamos que cada miembro de una población tiene asociado un valor numérico. Para que la muestra nos permita hacer inferencias sobre determinados parámetros de la población total será necesario asumir ciertas hipótesis sobre los valores de la población y sobre la relación existente entre la muestra y la población. Una de esas hipótesis es que exista una distribución de probabilidad subyacente a los valores de la población. Esto es, se asume que los valores de diferentes miembros de la población son variables aleatorias independientes que siguen una misma distribución. En concreto, los datos de la muestra son variables aleatorias que tienen una misma distribución común. De esta forma, si observamos los datos de la muestra, seremos capaces de sacar conclusiones acerca de esta distribución poblacional subyacente.

Definición

Si $X_1, \ldots, X_n$ son variables aleatorias independientes siguiendo una misma distribución de probabilidad, se dice que constituyen una *muestra* procedente de dicha distribución.

En la mayor parte de las aplicaciones, la distribución poblacional no será completamente conocida, y se intentará utilizar la muestra para hacer inferencias sobre ella. Por ejemplo, un productor puede estar fabricando un nuevo tipo de baterías de coches con motores eléctricos. Estas baterías durarán un número aleatorio de millas siguiendo una distribución de probabilidad desconocida. Para averiguar cuál es la distribución de probabilidad subyacente, el productor puede fabricar y probar en carretera un determinado conjunto de baterías. Los datos resultantes, referidos al número de millas recorridas con cada batería, constituirán una muestra extraída de dicha distribución.

En este capítulo estaremos interesados en obtener las distribuciones de probabilidad de ciertos estadísticos que aparecen en los procesos de muestreo, entendiéndose por estadístico una magnitud numérica cuyo valor viene determinado por la muestra. Dos estadísticos importantes que se considerarán son la media muestral y la varianza muestral. En la sección 7.3 se analizará la media muestral y se obtendrán la esperanza y la varianza de este estadístico. Se verá que, cuando el tamaño muestral es relativamente grande, la distribución de probabilidad de la media muestral puede aproximarse por una distribución normal. Este hecho, que se deriva de uno de los resultados más importantes de la teoría de la probabilidad, conocido como *teorema central del límite*, será analizado en la sección 7.4. En la sección 7.5 se estudiarán situaciones en las que las muestras se extraen de una población finita de objetos, y se explicará qué significa que una muestra sea *aleatoria*. Por lo general, cuando el tamaño de la población es grande en relación con el tamaño muestral, se tratará la población como si fuera infinita. Se comentará y se explicará exactamente cuándo se puede hacer esto y qué consecuencias entraña. En la sección 7.6 se considerará la distribución de la varianza muestral cuando la muestra procede de una población normal.

7.3 Media Muestral

Consideremos una población en la que cada uno de sus elementos tiene asignado un valor numérico. Por ejemplo, la población podría estar formada por los miembros adultos de una determinada comunidad, y el valor asignado a cada adulto podría ser su renta anual, o su altura, o su edad, etcétera. Por lo general se supondrá que el valor asociado a cada miembro de la población se puede considerar como el valor de una variable aleatoria con esperanza μ y varianza σ^2. Los valores μ y σ^2 se denominarán *media poblacional* y *varianza poblacional*, respectivamente. Sean $X_1, X_2, \ldots, X_n$ los valores de una muestra extraída de esa población. La media muestral se define como

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n}$$

Puesto que el valor de la media muestral $\overline{X}$ viene determinado por los valores de las variables aleatorias de la muestra, se tendrá que $\overline{X}$ también será una variable aleatoria. Se puede demostrar que su esperanza es

$$E[\overline{X}] = \mu$$

Es decir, el valor esperado de la media muestral $\overline{X}$ es igual a la media poblacional μ.

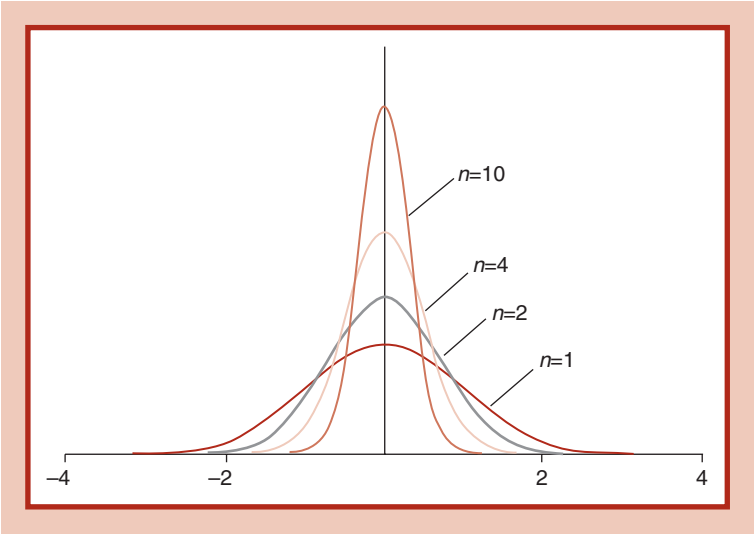


Figura 7.1 Densidades de las medias muestrales procedentes de una población normal estándar.

También se puede demostrar que la varianza de la media muestral es

$$\text{Var}(\overline{X}) = \frac{\sigma^2}{n}$$

Así pues, se ve que la media muestral $\overline{X}$ tiene el mismo valor esperado que cualquier valor de dato individual, mientras que su varianza es menor que la de un valor de dato individual, al venir multiplicada por el factor $1/n$, siendo n el tamaño de la muestra. Se puede concluir, pues, que $\overline{X}$ está centrada sobre la media poblacional μ, pero su dispersión disminuye más y más a medida que el tamaño muestral aumenta. La figura 7.1 representa gráficamente las funciones de densidad de probabilidad de la media muestral para distintos tamaños muestrales, cuando se muestrea sobre una población normal estándar.

Ejemplo 7.1 Comprobemos las anteriores fórmulas de la esperanza y la varianza de la media muestral si se consideran muestras de tamaño 2 procedentes de una población cuyos valores tienen la misma probabilidad de ser 1 o 2. Es decir, si X es el valor de un miembro de la población, se tendrá que

$$P\{X = 1\} = \frac{1}{2}$$

$$P\{X = 2\} = \frac{1}{2}$$

La media y la varianza de la población se obtienen como sigue:

$$\mu = E[X] = 1\left(\frac{1}{2}\right) + 2\left(\frac{1}{2}\right) = 1{,}5$$

y

$$\begin{aligned}\sigma^2 = \text{Var}(X) &= E[(X - \mu)^2] \\ &= (1 - 1{,}5)^2\left(\frac{1}{2}\right) + (2 - 1{,}5)^2\left(\frac{1}{2}\right) \\ &= \frac{1}{4}\end{aligned}$$

Para obtener la distribución de probabilidad de la media muestral $(X_1 + X_2)/2$, observe que el par X_1, X_2 puede tomar cualquiera de los cuatro pares de valores posibles

$$(1, 1), (1, 2), (2, 1), (2, 2)$$

donde, por ejemplo, el par (2, 1) significa que $X_1 = 2$ y $X_2 = 1$. De la independencia de X_1 y X_2 se desprende que la probabilidad de cualquiera de los cuatro pares de datos es 1/4. Así pues, se ve que los posibles valores de $\overline{X} = (X_1 + X_2)/2$ junto con sus probabilidades asociadas, son los siguientes:

$$P\{\overline{X} = 1\} = P\{(1, 1)\} = \frac{1}{4}$$

$$P\{\overline{X} = 1{,}5\} = P\{(1, 2) \text{ o } (2, 1)\} = \frac{2}{4} = \frac{1}{2}$$

$$P\{\overline{X} = 2\} = P\{(2, 2)\} = \frac{1}{4}$$

Por consiguiente,

$$E[\overline{X}] = 1\left(\frac{1}{4}\right) + 1{,}5\left(\frac{1}{2}\right) + 2\left(\frac{1}{4}\right) = \frac{6}{4} = 1{,}5$$

y

$$\begin{aligned}\text{Var}(\overline{X}) &= E[(\overline{X} - 1{,}5)^2] \\ &= (1 - 1{,}5)^2\left(\frac{1}{4}\right) + (1{,}5 - 1{,}5)^2\left(\frac{1}{2}\right) + (2 - 1{,}5)^2\left(\frac{1}{4}\right) \\ &= \frac{1}{16} + 0 + \frac{1}{16} = \frac{1}{8}\end{aligned}$$

con lo cual, dado que $\mu = 1{,}5$ y $\sigma^2 = 1/4$, queda efectivamente comprobado que $E[\overline{X}] = \mu$ y que $\text{Var}(\overline{X}) = \sigma^2/2$.

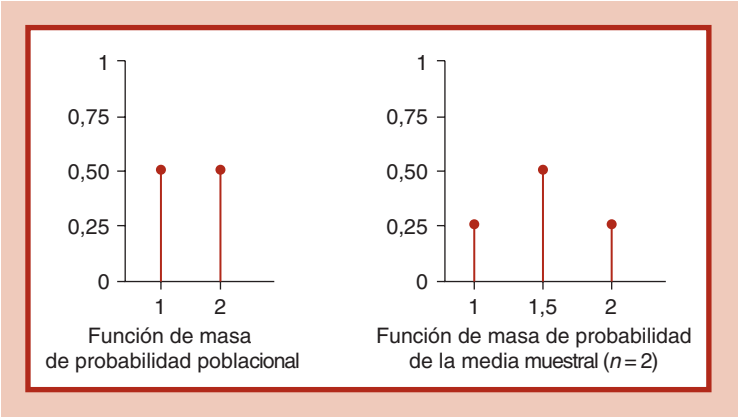


Figura 7.2 Funciones de masa de probabilidad.

La figura 7.2 representa gráficamente la distribución de probabilidad poblacional junto con la distribución de probabilidad de la media muestral de la muestra de tamaño 2. ■

La desviación típica de una variable aleatoria, que coincide con la raíz cuadrada de su varianza, es un indicador directo de la dispersión de su distribución. Se deduce de la igualdad

$$\text{Var}(\overline{X}) = \frac{\sigma^2}{n}$$

que $\text{SD}(\overline{X})$, la desviación típica de la media muestral $\overline{X}$, viene dada por

$$\text{SD}(\overline{X}) = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

En la expresión anterior, σ es la desviación típica de la población, y n el tamaño muestral.

La *desviación típica* de la media muestral es igual a la desviación típica de la población dividida por la raíz cuadrada del tamaño muestral.

Resumiendo, en este apartado se ha visto que para una muestra de tamaño n la esperanza de la media muestral es igual a la media poblacional y que la varianza de la media muestral es igual a la varianza de la población multiplicada por el factor reductor $1/n$. Ahora bien, aunque el conocimiento de la media y la varianza de un estadístico indique algo acerca de su distribución de probabilidad, aún queda mucho por determinar. Se verá en la sección 7.4 que la distribución de probabilidad de la media muestral es aproximadamente normal y,

como se sabe, la distribución normal queda completamente determinada si se conoce su media y su varianza.

Problemas

1. Considere la población descrita en el ejemplo 7.1. Represente gráficamente los posibles valores de la media muestral, junto con sus probabilidades asociadas, cuando el tamaño muestral es:

   (a) $n = 3$

   (b) $n = 4$

   En ambos casos, obtenga también la desviación típica de la media muestral.

2. Supongamos que X_1 y X_2 constituyen una muestra de tamaño 2 procedente de una población en la que un miembro típico X puede tomar los valores 1 o 2 con probabilidades:

$$P\{X = 1\} = 0{,}7 \qquad P\{X = 2\} = 0{,}3$$

   (a) Calcule $E[X]$.

   (b) Obtenga $\mathrm{Var}(X)$.

   (c) ¿Cuáles son los posibles valores de $\overline{X} = (X_1 + X_2)/2$?

   (d) Determine las probabilidades de que $\overline{X}$ tome cada uno de los valores obtenidos en (c).

   (e) Con los resultados del apartado (d), obtenga directamente $E[\overline{X}]$ y $\mathrm{Var}(\overline{X})$.

   (f) ¿Los resultados obtenidos en (a), (b) y (e) son coherentes con las fórmulas introducidas en esta sección?

3. Consideremos una población cuyos valores y probabilidades asociadas vienen dados por

$$p(1) = p(2) = p(3) = \frac{1}{3}$$

   (a) Obtenga $E[X]$.

   (b) Calcule $\mathrm{SD}(X)$.

   (c) Sea $\overline{X}$ la media muestral de una muestra de tamaño 2 extraída de esa población. Determine todos los valores posibles de $\overline{X}$ junto con sus probabilidades asociadas.

   (d) Utilice los resultados del apartado (c) para calcular $E[\overline{X}]$ y $\mathrm{SD}(\overline{X})$.

   (e) ¿Los resultados obtenidos son coherentes?

4. La cantidad de dinero retirada de un cajero automático de una sucursal del Banco de América en cada transacción tiene media 80 $ y desviación típica 80 $. ¿Cuáles son la media y la desviación típica de la cantidad media retirada en las 20 próximas transacciones?

5. Un productor de tabaco mantiene que el contenido medio de nicotina de sus cigarrillos es de 2,4 miligramos con una desviación típica de 0,2 miligramos. Si se acepta que estas cifras son correctas, calcule el valor esperado y la varianza del contenido medio muestral en nicotina de

   (a) 36 (b) 64 (c) 100 (d) 900

   cigarrillos seleccionados aleatoriamente.

6. Cierto tipo de bombillas eléctricas tienen una duración con un valor esperado de 475 horas y una desviación típica de 60 horas. Calcule el valor esperado y la desviación típica de la media muestral de

   (a) 100 (b) 200 (c) 400

   bombillas.

7. El peso de una persona elegida aleatoriamente entre los viajeros de un trasbordador tiene un valor esperado de 155 libras y una desviación típica de 28 libras. El trasbordador tiene una capacidad para llevar a 100 viajeros. Calcule la esperanza y la desviación típica del peso total de los pasajeros del trasbordador si éste va completo.

7.4 Teorema central del límite

En el apartado anterior se vio que, si se extrae una muestra de tamaño n de una población cuyos elementos tienen media μ y desviación típica σ, la media muestral $\overline{X}$ tiene media μ y desviación típica $\sigma\sqrt{n}$. En esta sección se presentará uno de los más importantes resultados de la teoría de la probabilidad, conocido como el *teorema central del límite*, que establece que la suma (y, por consiguiente, también la media) de un gran número de variables aleatorias independientes sigue aproximadamente una distribución normal.

Teorema central del límite

Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria procedente de una población con media μ y desviación típica σ. Si n es suficientemente grande, la suma

$$X_1 + X_2 + \cdots + X_n$$

sigue aproximadamente una distribución normal con media μ y desviación típica $\sigma\sqrt{n}$.

Ejemplo 7.2 Una compañía aseguradora de automóviles tiene 10 000 ($= 10^4$) asegurados. Si el gasto anual que un asegurado ocasiona a la compañía tiene por media 260 dólares con una desviación típica de 800 dólares, aproxime la probabilidad de que el gasto total que la compañía debe afrontar en un año sobrepase 2,8 millones ($= 2{,}8 \times 10^6$) de dólares.

Solución Numeremos a los asegurados y sea X_i el gasto que ocasiona a la compañía el asegurado i, $i = 1, \ldots, 10^4$. Por el teorema central del límite, $X = \sum_{i=1}^{10^4} X_i$ sigue aproximada-

mente una distribución normal con media $10^4 \times 260 = 2{,}6 \times 10^6$ y desviación típica $800\sqrt{10^4} = 800 \times 10^2 = 8 \times 10^4$. Por consiguiente,

$$
\begin{aligned}
P\{X > 2{,}8 \times 10^6\} &= P\left\{\frac{X - 2{,}6 \times 10^6}{8 \times 10^4} > \frac{2{,}8 \times 10^6 - 2{,}6 \times 10^6}{8 \times 10^4}\right\} \\
&\approx P\left\{Z > \frac{0{,}2 \times 10^6}{8 \times 10^4}\right\} \\
&= P\left\{Z > \frac{20}{8}\right\} \\
&= P\{Z > 2{,}5\} = 0{,}0062
\end{aligned}
$$

donde $\approx$ significa "es aproximadamente igual a". Esto es, existen 6 posibilidades sobre 1000 de que el coste anual total que debe afrontar la compañía sobrepase 2,8 millones de dólares. ■

La anterior versión del teorema central del límite no es la más general, puesto que se puede demostrar que $\sum_{i=1}^{n} X_i$ sigue aproximadamente una distribución normal incluso aunque las variables aleatorias X_i sigan distribuciones distintas. De hecho, si todas las variables aleatorias tienden a ser, más o menos, de la misma magnitud, de forma que ninguna de ellas domine el valor de la suma, se puede demostrar que la suma de un gran número de variables aleatorias independientes sigue aproximadamente una distribución normal.

El teorema central del límite no nos proporciona únicamente un método para calcular la distribución de una suma de variables aleatorias, sino que además nos ayuda a explicar el hecho observable de que las frecuencias empíricas de un gran número de poblaciones existentes en la naturaleza exhiban una forma acampanada (es decir, normal). Verdaderamente, una de las primeras consecuencias del teorema central del límite permitió encontrar una justificación teórica al hecho empírico de que los errores de medida tendían a estar distribuidos normalmente. Esto es, si se considera que un error de medición es la consecuencia de un gran número de pequeños errores independientes, el teorema central del límite implica que su distribución será aproximadamente normal. Por ejemplo, se podría considerar que el error cometido en una medición astronómica es consecuencia de la suma de pequeños errores causados por hechos tales como:

1. los efectos de la temperatura sobre el aparato de medida,
2. el doblamiento del aparato debido a los rayos del sol,
3. los efectos elásticos,
4. las corrientes de aire,
5. las vibraciones del aire y
6. los errores humanos.

En consecuencia, por el teorema central del límite, el error total de medición seguirá aproximadamente una distribución normal. De ahí se deriva que el histograma de los errores resultantes de una serie de medidas repetidas de un *mismo* objeto tienda a tener la forma acampanada propia de la distribución normal.

El teorema central del límite también permite explicar parcialmente por qué la distribución de una gran cantidad de conjuntos de datos relacionados con distintas características biológicas tiende a ser aproximadamente normal. Por ejemplo, consideremos a una pareja particular, que llamaremos María y Pedro Fontánez, y observemos las alturas de sus hijas (por ejemplo, al cumplir 20 años). Se podría pensar que la altura de una hija determinada es consecuencia de la suma de un gran número de variables aleatorias independientes, que se deben, entre otras causas, al conjunto aleatorio de genes que recibe de sus padres así como a factores ambientales. Dado que cada una de esas variables desempeña un pequeño papel sobre la altura resultante, parece razonable pensar que, a raíz del teorema central del límite, la altura de una hija siga una distribución normal. Si la familia Fontánez tuviera muchas hijas, el histograma de las alturas de todas ellas reflejaría a grandes rasgos la forma de la curva normal. (Lo mismo ocurriría con los hijos de Pedro y María, aunque la curva normal de los hijos podría tener unos parámetros distintos de los de las hijas. Esto es, no se podría utilizar el teorema central del límite para concluir que la altura de todos los descendientes de los Fontánez debe seguir una distribución normal, si el factor sexo no desempeñara un papel "pequeño" en la determinación de las alturas.)

Así pues, se puede utilizar el teorema central del límite para explicar por qué las alturas de todas las hijas posibles de una determinada pareja siguen una curva normal. Sin embargo, por sí mismo el teorema no explica por qué en un histograma las alturas de un conjunto de hijas de padres diferentes siguen también una distribución normal. Para razonar por qué no, supongamos que dicho conjunto de hijas incluye tanto a una hija de María y Pedro Fontánez como a una hija de Enrique y Catalina Silva. Por el mismo argumento empleado anteriormente, la altura de la hija de los Silva se distribuirá normalmente, al igual que ocurría con la hija de los Fontánez. Sin embargo, los parámetros de las dos distribuciones normales –una por cada familia– no tienen por qué coincidir. (Por ejemplo, si Catalina y Enrique miden ambos 6 pies de altura, mientras que María y Pedro miden los dos alrededor de 5 pies, parecería claro que las alturas de sus hijas podrían seguir distribuciones normales distintas.) Por el mismo razonamiento, se podría concluir que las alturas de un conjunto grande de mujeres de diferentes familias podrían provenir de distintas distribuciones normales. En consecuencia, no resulta sencillo explicar que el gráfico de todas estas alturas tenga por qué seguir una distribución normal. (En el capítulo 12 se dará una explicación más completa de por qué los conjuntos de datos biológicos siguen habitualmente una distribución normal.)

7.4.1 Distribución de la media muestral

Se puede utilizar el teorema central del límite para calcular la distribución de la media muestral. Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria procedente de una población con media μ y varianza σ^2, y sea

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Perspectiva histórica

La aplicación del teorema central del límite como justificación de que los errores de medida se distribuyen aproximadamente según una normal es considerada como una de las más importantes aportaciones científicas. En realidad, en los siglos XVII y XVIII, el teorema central del límite se conoció como la *ley de frecuencias de los errores.*

Muchos científicos pensaron que la *ley de frecuencias de los errores* suponía un gran avance. Observe, al respecto, el comentario de Francis Galton (de su libro *La herencia natural*, publicado en 1889):

Difícilmente conozco algo que alimente tanto mi imaginación como el maravilloso orden cósmico que se deriva de la "Ley de frecuencias de los errores". Si los griegos hubieran conocido esta ley, seguro que la habrían endiosado. Reina con serenidad y en completa auto-modestia entre la confusión más salvaje. Cuanto más vigentes están la ley de la calle y la aparente anarquía, más perfecto es su balanceo. Es la ley suprema de la sinrazón.

la media muestral. Puesto que la multiplicación de una normal por una constante continúa siendo normal, se sigue del teorema central del límite que $\overline{X}$ (que es igual a $\sum_{i=1}^{n} X_i$ multiplicada por la constante $1/n$) se distribuirá aproximadamente como una normal, si el tamaño muestral es grande. Si se tiene en cuenta que $\overline{X}$ tiene media μ y desviación típica $\sigma/\sqrt{n}$, la variable estandarizada

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

seguirá aproximadamente una distribución normal estándar.

Sea $\overline{X}$ la media muestral de una muestra de tamaño n procedente de una población con media μ y varianza σ^2. Por el teorema central del límite,

$$P\{\overline{X} \le a\} = P\left\{\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \le \frac{a - \mu}{\sigma/\sqrt{n}}\right\}$$

$$\approx P\left\{Z \le \frac{a - \mu}{\sigma/\sqrt{n}}\right\}$$

donde Z representa una normal estándar.

Ejemplo 7.3 Los niveles de colesterol en la sangre de una población de trabajadores tiene media 202 y desviación típica 14.

(a) Si se selecciona una muestra de 36 trabajadores, aproxime la probabilidad de que la media muestral de sus niveles de colesterol esté comprendida entre 198 y 206.

(b) Repita el apartado (a) para un tamaño muestral igual a 64.

Solución

(a) Se deduce del teorema central del límite que $\overline{X}$ se distribuye aproximadamente como una normal de media $\mu = 202$ y desviación típica $\sigma/\sqrt{n} = 14/\sqrt{36} = 7/3$. Así pues, la variable estandarizada

$$W = \frac{\overline{X} - 202}{7/3}$$

tiene aproximadamente la distribución de una normal estándar. Para calcular $P\{198 \leq \overline{X} \leq 206\}$, escribiremos en primer lugar la desigualdad en términos de la variable estandarizada W. Esto nos lleva a escribir

$$\begin{aligned} P\{198 \leq \overline{X} \leq 206\} &= P\left\{\frac{198 - 202}{7/3} \leq \frac{\overline{X} - 202}{7/3} \leq \frac{206 - 202}{7/3}\right\} \\ &= P\{-1{,}714 \leq W \leq 1{,}714\} \\ &\approx P\{-1{,}714 \leq Z \leq 1{,}714\} \end{aligned}$$

donde Z es una variable aleatoria normal estándar,

$$\begin{aligned} &= 2P\{Z \leq 1{,}714\} - 1 \\ &= 0{,}913 \end{aligned}$$

y la última igualdad se obtiene a partir de la tabla D.1 del Apéndice D (o del Programa 6-1).

(b) Para una muestra de tamaño 64, la media muestral $\overline{X}$ tendrá media 202 y desviación típica $14/\sqrt{64} = 7/4$. Por consiguiente, si se escribe la probabilidad pedida en términos de la variable estandarizada

$$\frac{\overline{X} - 202}{7/4}$$

se obtiene

$$\begin{aligned} P\{198 \leq \overline{X} \leq 206\} &= P\left\{\frac{198 - 202}{7/4} \leq \frac{\overline{X} - 202}{7/4} \leq \frac{206 - 202}{7/4}\right\} \\ &\approx P\{-2{,}286 \leq Z \leq 2{,}286\} \\ &= 2P\{Z \leq 2{,}286\} - 1 \\ &= 0{,}978 \end{aligned}$$

Así pues, cuando se aumenta el tamaño muestral de 36 a 64, se ve que también aumenta, de 0,913 a 0,978, la probabilidad de que la media muestral difiera de la media poblacional en menos de 4 unidades. ■

Ejemplo 7.4 Una astrónoma pretende medir, en unidades de años luz, la distancia existente entre su laboratorio y una estrella muy alejada. Sin embargo, la astrónoma sabe que, debido a las distintas condiciones atmosféricas y a los errores normales, cada vez que realiza una medida no obtiene la distancia exacta, sino sólo una estimación de ella. Como resultado, ha decidido realizar 10 mediciones distintas y utilizar la media de todas ellas como un valor estimado de la distancia real. Si los valores de las 10 mediciones realizadas constituyen una muestra extraída de la población de todas las posibles mediciones y esta población tiene media d (la distancia real) y desviación típica 3 años luz, calcule la probabilidad de que el valor estimado por la astrónoma difiera de la distancia real en menos de 0,5 años luz.

Solución La probabilidad pedida es

$$P\{-0{,}5 < \overline{X} - d < 0{,}5\}$$

donde $\overline{X}$ representa la media muestral de las 10 mediciones. Puesto que $\overline{X}$ tiene media d y desviación típica $3/\sqrt{10}$, esta probabilidad se debería escribir en términos de la variable estandarizada

$$\frac{\overline{X} - d}{3/\sqrt{10}}$$

Esto conduce a

$$\begin{aligned} P\{-0{,}5 < \overline{X} - d < 0{,}5\} &= P\left\{\frac{-0{,}5}{3/\sqrt{10}} < \frac{\overline{X} - d}{3/\sqrt{10}} < \frac{0{,}5}{3/\sqrt{10}}\right\} \\ &\approx P\left\{\frac{-0{,}5}{3/\sqrt{10}} < Z < \frac{0{,}5}{3/\sqrt{10}}\right\} \\ &= P\{-0{,}527 < Z < 0{,}527\} \\ &= 2P\{Z < 0{,}527\} - 1 = 0{,}402 \end{aligned}$$

Así pues, se ve que con 10 mediciones existe una probabilidad del 40,2% de que la distancia estimada difiera de la distancia real en más o menos 0,5 años luz como máximo. ■

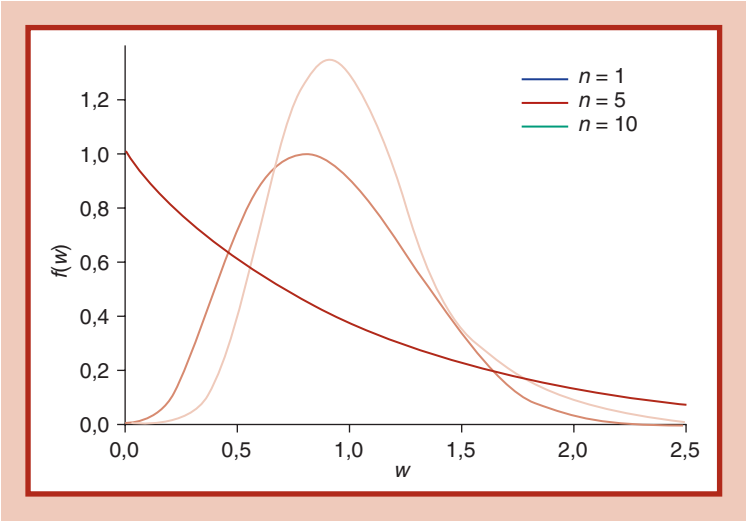


Figura 7.3 Densidad de la media de n variables aleatorias exponenciales.

7.4.2 ¿Cómo debe ser de grande la muestra?

El teorema central del límite deja abierta la cuestión de cuán grande debe ser el tamaño de la muestra n para que la aproximación normal sea válida; la respuesta depende de la distribución de la población subyacente a los datos muestrales. Por ejemplo, si dicha distribución subyacente es normal, la media muestral $\overline{X}$ es siempre normal, independientemente del tamaño muestral. Una regla empírica es que, por lo general, se puede utilizar la aproximación normal siempre que el tamaño muestral sea como mínimo 30. En términos prácticos, esto significa que, sin importarnos en qué medida la distribución de la población subyacente difiere de la normalidad, la media muestral de una muestra de tamaño 30, o más, siempre se puede aproximar por la normal. En la mayoría de los casos, sin embargo, la aproximación normal es válida para tamaños muestrales mucho más reducidos. En la figura 7.3 se reflejan las distribuciones de las medias muestrales para tamaños muestrales $n = 1$, 5 y 10 y para una determinada distribución de la población subyacente (conocida como *distribución exponencial*).

Problemas

1. Consideremos una muestra extraída de una población con media 128 y desviación típica 16. Calcule la probabilidad aproximada de que la media muestral esté comprendida entre 124 y 132, cuando el tamaño muestral es:

   (a) $n = 9$

   (b) $n = 25$

   (c) $n = 100$

(Culver)

Pierre Simon, Marqués de Laplace

Perspectiva histórica

El teorema central del límite fue inicialmente formulado y demostrado por el matemático francés Pierre Simon, Marqués de Laplace, tras haber observado que los errores de medidas (que se pueden considerar como el resultado de un gran número de causas ligeras) tienden a distribuirse normalmente. Laplace, que también fue un famoso astrónomo (de hecho, se le llegó a llamar "el Newton francés"), fue uno de los científicos que inicialmente más contribuyeron al desarrollo de la Probabilidad y la Estadística. Llegó a popularizar el uso de la probabilidad en los problemas de la vida cotidiana. Estaba totalmente convencido de la importancia de esta ciencia, como se pone de manifiesto en la siguiente cita, extraída de su libro *Teoría analítica de la Probabilidad.*

> Vemos que la Teoría de la Probabilidad en el fondo sólo es sentido común reducido a cálculo; nos hace apreciar con exactitud lo que las mentes razonables toman por un tipo de instinto, incluso sin ser capaces de darse cuenta . . . Es sorprendente que esta ciencia, que surgió del análisis de los juegos de azar, llegara a ser el objeto más importante del conocimiento humano. . . Las principales cuestiones de la vida son, en gran medida, meros problemas de probabilidad.

Un apunte interesante sobre el teorema central del límite es que, debido a él, la mayor parte de los científicos de finales del siglo XIX y principios del XX creían que casi todos los conjuntos de datos eran normales. En palabras del famoso científico francés Henri Poincaré:

> Todo el mundo cree en él: los experimentalistas consideran que es un teorema matemático, los matemáticos piensan que se trata de un hecho empírico.

2. Los pasajeros habituales de una determinada línea aérea vuelan una cantidad aleatoria de millas al año, con media y desviación típica (en miles de millas) 23 y 11, respectivamente. Con carácter promocional, la línea aérea ha decidido seleccionar a 20 de esos pasajeros y darles, como premio, un cheque de 10 dólares por cada 1000 millas de vuelo. Calcule la probabilidad de que la cantidad total pagada por la compañía aérea como premio:

   (a) Esté comprendida entre 4500 y 5000 $.

   (b) Sea mayor que 5200 $.

3. En el ejemplo 7.2 calcule la probabilidad de que el pago anual al que debe hacer frente la compañía de seguros esté comprendido entre 2,5 y 2,7 millones de dólares.

4. Si un jugador apuesta 1 dólar a un número de la ruleta, puede ganar 35 dólares con una probabilidad de 1/38 o perder el dólar apostado con una probabilidad de 37/38. Denote por X a la ganancia obtenida en una apuesta de este tipo.

   (a) Obtenga $E[X]$ y SD(X).

Supongamos que el jugador decide realizar varias apuestas sucesivas. Demuestre que:

(b) La probabilidad de que vaya ganando tras 1000 apuestas es aproximadamente 0,39.

(c) La probabilidad de que vaya ganando tras 100 000 apuestas es aproximadamente 0,002.

5. El tiempo que se tarda en imprimir una fotografía es una variable aleatoria con media 17 segundos y desviación típica 0,8 segundos. Calcule la probabilidad de que se tarde en imprimir 1000 fotografías:

   (a) más de 7200 segundos

   (b) entre 1690 y 1710 segundos

6. Un semiconductor de circón es un elemento esencial para el funcionamiento de un superconductor y debe ser reemplazado de inmediato en cuanto falle. Su duración esperada es de 100 horas, y su desviación típica de 34 horas. Si se dispone de 22 de dichos semiconductores, calcule la probabilidad de que el superconductor esté operativo durante las próximas 2000 horas. (Esto es, calcule la probabilidad de que la suma de las duraciones de los 22 semiconductores supere las 2000 horas.)

7. La cantidad de papel que se emplea en un comercio de reprografía por cada trabajo encargado es, en media, 200 hojas, con desviación típica 50 hojas. Si se tienen a mano 2300 hojas y se deben reproducir 10 trabajos, ¿cuál es la probabilidad de que se puedan hacer las 10 reproducciones con el papel disponible?

8. Los servicios de mantenimiento de una autopista disponen de una cantidad suficiente de sal para poder afrontar un total de 80 pulgadas de nieve. Supongamos que la cantidad de nieve que cae diariamente tiene media 1,5 pulgadas y desviación típica 0,3 pulgadas.

   (a) Aproxime la probabilidad de que la sal disponible sea suficiente para hacer frente a las nevadas de los próximos 50 días.

   (b) ¿Qué hipótesis se debe asumir para resolver el apartado (a)?

   (c) ¿Es razonable esta hipótesis? ¡Explíquelo brevemente!

9. Se redondean 50 números al entero más próximo y después se suman. Si los errores de redondeo individuales se distribuyen uniformemente entre –0,5 y 0,5, ¿cuál es la probabilidad aproximada de que la suma resultante difiera de la suma exacta en más de 3 unidades? (Utilice el hecho de que la media y la varianza de una variable aleatoria distribuida uniformemente entre –0,5 y 0,5 son 0 y 1/12, respectivamente.)

10. Se lanza sucesivamente un dado, con seis caras igualmente probables, hasta que la suma de los resultados obtenidos sea mayor que 400. ¿Cuál es la probabilidad aproximada de que el número de lanzamientos realizados sea mayor que 140? (*Sugerencia*: Relacione esta probabilidad con la de que la suma de los 140 primeros sea menor que 400.)

11. En el ejemplo 7.4 calcule la probabilidad de que la estimación de la astrónoma difiera de la distancia real en menos de 0,5 años luz, si:

    (a) Realiza un total de 100 observaciones.

    (b) Lleva a cabo 10 observaciones tras descubrir un método para mejorar las mediciones, de forma que la desviación típica de cada observación se reduzca de 3 a 2 años luz.

12. Supongamos que las baterías de un coche eléctrico funcionan durante un número de millas que tiene media m y desviación típica 100. Mediante el teorema central del límite, aproxime la probabilidad de que el número medio de millas recorridas por batería, para un conjunto de n de éstas, difiera de μ en más de 20 si:

    (a) $n = 10$ (b) $n = 20$ (c) $n = 40$ (d) $n = 100$

13. Un productor de tabaco mantiene que el contenido medio de nicotina en sus cigarrillos es de 2,4 miligramos, con una desviación típica de 0,2 miligramos. Si se asume que estas cifras son correctas, calcule la probabilidad de que la media muestral de 100 cigarrillos seleccionados aleatoriamente sea:

    (a) mayor que 2,5 miligramos

    (b) menor que 2,25 miligramos

14. La duración esperada de un determinado tipo de bombilla eléctrica es de 500 horas, con una desviación típica de 60 horas. Calcule la probabilidad de que la media muestral de las duraciones de 20 bombillas sea menor que 480 horas.

15. Consideremos una muestra de tamaño 16 procedente de una población con media 100 ydesviación típica σ. Calcule la probabilidad de que la media muestral esté comprendida entre 96 y 104 cuando:

    (a) $\sigma = 16$ (b) $\sigma = 8$ (c) $\sigma = 4$

    (d) $\sigma = 2$ (b) $\sigma = 1$

16. Por su experiencia, un profesor sabe que las calificaciones de los estudiantes tienen media 77 y desviación típica 15. En la actualidad, el profesor imparte clases en dos cursos distintos: uno de tamaño 25 y el otro de tamaño 64.

    (a) Calcule la probabilidad de que la calificación media del curso con 25 estudiantes esté comprendida entre 72 y 82 puntos.

    (b) Repita el apartado (a) para el curso con 64 estudiantes.

    (c) ¿Cuál es la probabilidad aproximada de que la calificación media de la clase de tamaño 25 sea mayor que la de la clase de tamaño 64?

    (d) Supongamos que las calificaciones medias de las dos clases son 76 y 83. ¿Cuál de las dos clases –la de 25 o la de 64 alumnos– parece tener mayor probabilidad de ser la que obtuvo la calificación media de 83 puntos? ¡Explique la respuesta intuida!

7.5 Muestreo de proporciones en poblaciones finitas

Consideremos una población de tamaño N en la que determinados elementos presentan cierta característica de interés. Denotemos por p a la proporción de individuos de la población que muestran la característica. Así pues, Np elementos de la población presentan la característica y $N(1 - p)$ elementos no la presentan.

Ejemplo 7.5 Supongamos que 60 de un total de 900 alumnos de una determinada escuela son zurdos. Si el ser zurdo es la característica de interés, $N = 900$ y $p = 1/15$. ■

Una muestra de tamaño n se dice que es una *muestra aleatoria* si se selecciona de forma que todos los posibles subconjuntos de la población de tamaño n tengan la misma probabilidad de ser la muestra. Por ejemplo, si la población consta de tres elementos *a, b, c*, una muestra aleatoria de tamaño 2 es aquella en la que cualquier subconjunto $\{a, b\}$, $\{a, c\}$ y {b, c} tiene la misma probabilidad de ser elegido. Se puede seleccionar secuencialmente una muestra aleatoria si se elige, primero, un elemento de la población en la que todos tengan la misma probabilidad de ser seleccionados; después se elige un segundo elemento entre los $N - 1$ elementos de la población restantes en la que todos tengan la misma probabilidad; y así sucesivamente.

Definición

Una muestra de tamaño n, extraída de una población de N elementos, se dice que es una *muestra aleatoria* si se selecciona de tal forma que cualquier subconjunto de n elementos de la población tiene la misma probabilidad de coincidir con la muestra.

En el Apéndice C se explica la mecánica operativa para seleccionar una muestra aleatoria usando un ordenador. (Adicionalmente, para llevar a cabo esta tarea se puede utilizar el Programa A-1, incluido en el disco adjunto a este libro.)

Supongamos ahora que se ha seleccionado una muestra aleatoria de tamaño n. Para $i = 1, \ldots, n$, definamos

$$X_i = \begin{cases} 1 & \text{si el } i\text{-ésimo elemento de la muestra presenta la característica} \\ 0 & \text{en otro caso} \end{cases}$$

Consideremos la suma de las X_i; esto es, consideremos

$$X = X_1 + X_2 + \cdots + X_n$$

Puesto que el término X_i contribuye a la suma con 1 unidad si el i-ésimo miembro de la muestra presenta la característica y contribuye con 0 en otro caso, se tiene que la suma anterior es igual al número de elementos muestrales que poseen la característica. (Por ejemplo, supongamos que $n = 3$, $X_1 = 1$, $X_2 = 0$ y $X_3 = 1$. En este caso, los miembros 1 y 3 de la muestra poseen la característica, mientras que el miembro 2 no la presenta. Así pues, exactamente 2 de los miembros muestrales presentan la característica, tal como se indica en la suma $X_1 + X_2 + X_3 = 2$.) De la misma manera, la media muestral

$$\overline{X} = \frac{X}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$$

será igual a la *proporción* de elementos muestrales que presentan la característica. Consideremos ahora las probabilidades asociadas con el estadístico .

Dado que los N elementos de la población tienen la misma probabilidad de ser el i-ésimo miembro de la muestra y que existen Np individuos de la población que presentan la característica, se tiene que

$$P\{X_i = 1\} = \frac{Np}{N} = p$$

Además,

$$P\{X_i = 0\} = 1 - P\{X_i = 1\} = 1 - p$$

Es decir, cada X_i puede tomar los valores 1 o 0 con unas probabilidades respectivas p y $1 - p$.

Observe que las variables aleatorias $X_1, \ldots, X_n$ no son independientes. Por ejemplo, si se tiene en cuenta que el segundo miembro de la muestra puede ser uno cualquiera de la población, todos con la misma probabilidad, y que existen Np elementos poblacionales que presentan la característica, se tiene que la probabilidad de que el segundo elemento muestral presente la característica es $Np/N = p$. Esto es, sin saber nada sobre el resultado de la primera selección muestral, se tiene que

$$P\{X_2 = 1\} = p$$

Por el contrario, la probabilidad de que $X_2 = 1$ condicionada a que el primer elemento muestral presente la característica es

$$P\{X_2 = 1 | X_1 = 1\} = \frac{Np - 1}{N - 1}$$

sin más que tener en cuenta que, si el primer elemento de la muestra presenta la característica, cualquiera de los $N - 1$ elementos restantes de la población puede, con la misma probabilidad, ser el segundo miembro de la muestra, y $Np - 1$ de estos tienen la característica. De igual forma, la probabilidad de que el segundo elemento de la muestra presente la característica, dado que el primero no la presente, es

$$P\{X_2 = 1 | X_1 = 0\} = \frac{Np}{N - 1}$$

En consecuencia, conocer que el primer elemento muestral presenta la característica modifica las probabilidades de selección del segundo elemento. Pese a ello, si el tamaño poblacional N es grande en relación con el tamaño muestral n, la modificación resulta ser prácticamente irrelevante. Por ejemplo, si $N = 1000$ y $p = 0{,}4$, se tiene

$$P\{X_2 = 1 | X_1 = 1\} = \frac{399}{999} = 0{,}3994$$

que es un valor muy cercano a la probabilidad incondicional de que $X_2 = 1$, dada por

$$P\{X_2 = 1\} = 0{,}4$$

De la misma manera, la probabilidad de que el segundo elemento de la muestra presente la característica, supuesto que el primero no la presente, viene dada por

$$P\{X_2 = 1 | X_1 = 0\} = \frac{400}{999} = 0{,}4004$$

que, de nuevo, es muy próximo a 0,4.

En realidad, se puede demostrar que, cuando el tamaño de la población N es grande con respecto al tamaño muestral n, $X_1, \ldots, X_n$ son aproximadamente independientes. En consecuencia, si se visualiza cada X_i de forma que represente el resultado de una prueba de Bernoulli en la que el éxito se obtiene si X_i toma el valor 1 y el fracaso se obtiene en caso contrario, se tendrá que $\sum_{i=1}^{n} X_i$ se puede identificar con el número de éxitos obtenidos en las n pruebas. Si éstas fueran independientes, puesto que cada una de ellas tiene una probabilidad p de éxito, se tendría que X sería una variable aleatoria binomial de parámetros n y p.

En resumen, si X denota el número de elementos de la muestra que presentan la característica, se desprende de lo anterior que, si el tamaño poblacional es grande en relación con el tamaño muestral, la distribución de X sigue aproximadamente una binomial de parámetros n y p.

De aquí en adelante se supondrá que el tamaño de la población subyacente es grande en comparación con el tamaño muestral y, en consecuencia, se asumirá que la distribución de X es binomial.

Con los valores de la media y la desviación típica de las variables aleatorias binomiales obtenidas en la sección 5.5.1 se ve que

$$E[X] = np \quad \text{y} \quad \text{SD}(X) = \sqrt{np(1 - p)}$$

Dado que $\overline{X}$, la proporción de elementos muestrales que presentan la característica, es igual a X/n se ve que

$$E[\overline{X}] = \frac{E[X]}{n} = p$$

y

$$\text{SD}(\overline{X}) = \frac{\text{SD}(X)}{n} = \sqrt{\frac{p(1 - p)}{n}}$$

Ejemplo 7.6 Supongamos que el 50% de los elementos de una población piensa votar al candidato A en unas próximas elecciones. Si se extrae una muestra de tamaño 100, la proporción de miembros de la muestra que se declaran a favor de dicho candidato tendrá una media

$$E[\overline{X}] = 0{,}50$$

y una desviación típica

$$\text{SD}(\overline{X}) = \sqrt{\frac{0{,}50(1 - 0{,}50)}{100}} = \sqrt{\frac{1}{400}} = 0{,}05 \quad \blacksquare$$

7.5.1 Probabilidades asociadas a las proporciones muestrales: la aproximación normal a la distribución binomial

De nuevo denotemos como $\overline{X}$ a la proporción de elementos de una muestra aleatoria de tamaño n que presentan una determinada característica. Para obtener las probabilidades asociadas a la variable aleatoria $\overline{X}$ se hará uso del hecho de que $X = n\overline{X}$ sigue una binomial de parámetros n y p. Ahora bien, las probabilidades de las binomiales se pueden calcular mediante el teorema central del límite. De facto, desde un punto de vista histórico, una de las aplicaciones más importantes del teorema central del límite consistió en el cómputo de las probabilidades binomiales.

Para ver cómo se consigue esto, denotemos por X a una variable aleatoria binomial de parámetros n y p. Dado que X se puede identificar con el número de éxitos obtenidos en n pruebas independientes con probabilidad p de éxito, se puede escribir como

$$X = X_1 + X_2 + \cdots + X_n$$

donde

$$X_i = \begin{cases} 1 & \text{si en la prueba } i \text{ resulta un éxito} \\ 0 & \text{si en la prueba } i \text{ resulta un fracaso} \end{cases}$$

En los ejemplos 5.6 y 5.12 se vio que

$$E[X_i] = p \quad \text{y} \quad \text{Var}(X_i) = p(1 - p)$$

Se ve, pues, que X/n puede ser considerada como la media muestral de una muestra de tamaño n procedente de una población con media p y desviación típica $\sqrt{p(1 - p)}$. En consecuencia, se desprende del teorema central del límite que, para valores grandes de n,

$$\frac{X/n - p}{\sqrt{p(1 - p)/n}} = \frac{X - np}{\sqrt{np(1 - p)}}$$

seguirá aproximadamente una distribución normal estándar. (La figura 7.4 ilustra gráficamente cómo la distribución de probabilidad de una variable aleatoria binomial de parámetros n y p se aproxima más y más a la normal a medida que n crece.)

Desde un punto de vista práctico, la aproximación normal a la binomial es bastante buena siempre que n sea lo suficientemente grande como para que tanto np como $n(1 - p)$ sean mayores que 5.

Ejemplo 7.7 Supongamos que exactamente un 46% de la población está a favor de un determinado candidato. Si se extrae una muestra aleatoria de tamaño 200, ¿cuál es la probabilidad de que al menos 100 de ellos estén a favor del candidato?

Solución Si X es el número de elementos muestrales a favor del candidato, X será una variable aleatoria binomial con parámetros $n = 200$ y $p = 0{,}46$. La probabilidad pedida es $P\{X \geq 100\}$. Para utilizar la aproximación normal, observe en primer lugar que, puesto que

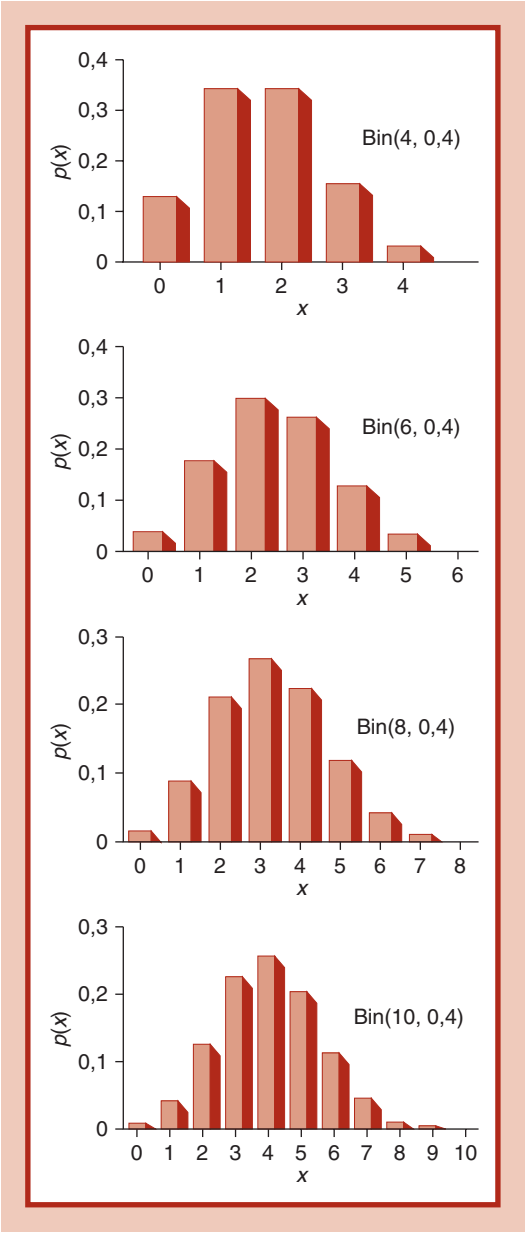


Figura 7.4 Las funciones de masa de probabilidad se aproximan más a la normal a medida que *n* crece.

la binomial es discreta mientras que la normal es una variable aleatoria continua, es mejor computar $P\{X = i\}$ como $P\{i - 0{,}5 \leq X \leq i + 0{,}5\}$ cuando se aplica la aproximación normal (esto se denomina *corrección por continuidad*). En consecuencia, para obtener $P\{X \geq 100\}$, se debería utilizar la aproximación normal sobre la probabilidad equivalente $P\{X \geq 99{,}5\}$ Si se considera la variable estandarizada

$$\frac{X - 200(0{,}46)}{\sqrt{200(0{,}46)(0{,}54)}} = \frac{X - 92}{7{,}0484}$$

se obtiene la siguiente aproximación normal a la probabilidad pedida:

$$\begin{aligned} P\{X \geq 100\} &= P\{X \geq 99{,}5\} \\ &= P\left\{\frac{X - 92}{7{,}0484} \geq \frac{99{,}5 - 92}{7{,}0484}\right\} \\ &\approx P\{Z > 1{,}0641\} \\ &= 0{,}144 \quad \text{(a partir de la tabla D.1 o del Programa 6-1)} \end{aligned}$$

El valor exacto de de la probabilidad pedida se puede obtener con el Programa 5-1. Con éste, la probabilidad exacta de que una binomial con parámetros $n = 200$ y $p = 0{,}46$ sea mayor o igual que 100 resulta ser 0,1437. En consecuencia, con la aproximación normal se consigue el valor correcto con tres cifras decimales. ■

Problemas

1. Supongamos que el 60% de los residentes de una ciudad están a favor de un cambio en el sistema de educación secundaria. Calcule la media y la desviación típica de la proporción de elementos de una muestra de tamaño n que estará a favor del cambio cuando:

   (a) $n = 10$ (b) $n = 100$

   (c) $n = 1000$ (d) $n = 10\,000$

2. El 10% de las baterías eléctricas de un grupo de ellas son defectuosas. Si se ha seleccionado aleatoriamente un conjunto de 8 de esas baterías, calcúlese la probabilidad de que:

   (a) Ninguna batería sea defectuosa.

   (b) Más de un 15% de las baterías sean defectuosas.

   (c) El número de baterías defectuosas esté comprendido ente 8 y 12.

3. Supongamos que en el problema 2 se han seleccionado aleatoriamente $n = 50$ baterías del grupo. Calcule las probabilidades de los apartados (a), (b) y (c) del problema anterior.

***Una historia aleccionadora: asegúrese de que se está muestreando sobre la población correcta**

La compañía X, a la que no se puede acceder mediante ningún tipo de transporte público y a la que sus empleados han de llegar en coche privado, tiene la intuición de que no existen suficientes acuerdos para compartir vehículos entre su personal. La compañía ha decidido que, si el número medio de trabajadores por coche es menor que 3, hará una propuesta para que sus empleados compartan coche y empezará a cobrar de forma inflexible una tarifa de aparcamiento a aquellos empleados que no compartan coche. Para analizar si dicha propuesta está justificada, se seleccionó aleatoriamente a 100 trabajadores a los que se les preguntó cuál era el número de trabajadores que viajaban en el coche con el que ellos habían accedido al trabajo ese día. La respuesta media fue 3,4; es decir, la suma de las 100 contestaciones divididas entre 100 fue igual a 3,4. En base a esto, la compañía decidió dejar las cosas como estaban y no realizar propuesta alguna. ¿Tomó la compañía una decisión correcta?

Esta pregunta tiene truco porque la compañía, al seleccionar una muestra aleatoria de 100 trabajadores, extrajo una muestra aleatoria de una población equivocada. Puesto que se pretendía conocer el número medio de trabajadores por coche, la compañía debería haber elegido una muestra aleatoria de la población de los coches aparcados –no de la población de trabajadores. Para ver por qué, consideremos un caso extremo en el que haya solamente 2 coches y 5 trabajadores, entre los que 4 de ellos comparten un coche y el otro no lo hace. Si se promediara sobre el número de coches se obtendría el valor $(4 + 1)/2 = 2{,}5$. Por el contrario, si se promediara sobre el número de trabajadores, teniendo en cuenta que 4 de los 5 trabajadores van en un coche con 4 pasajeros y el otro va solo, se obtendría como promedio $(4 + 4 + 4 + 4 + 1)/5 = 3{,}4$.

Se observa, pues, que la selección aleatoria de trabajadores (en lugar que de coches) produce que los coches con más pasajeros tienden a tener un mayor peso (debido a su mayor número de viajeros) en la muestra que el peso de los coches con menos pasajeros. Como resultado, el número medio de viajeros en los coches de los trabajadores seleccionados aleatoriamente tiende a ser mayor que el número medio de trabajadores por coche.

Para poder obtener un estimador correcto del número medio de trabajadores por coche, se debería haber extraído una muestra aleatoria de los coches del aparcamiento para después averiguar cuántos pasajeros ocuparon cada uno de ellos.

Debido a que la extracción de la muestra se llevó a cabo de manera equivocada, la compañía no puede concluir que el número medio de trabajadores por coche es como mínimo 3. De hecho, la compañía tendría que haber extraído una nueva muestra de coches en la forma correcta antes de decidir que las cosas siguieran tal como estaban.

4. Consideremos el problema 1. Calcule la probabilidad de que más de un 55% de los miembros de la muestra esté a favor de la propuesta si el tamaño muestral es:

   (a) $n = 10$

   (b) $n = 100$

(c) $n = 1000$

(d) $n = 10\,000$

La tabla siguiente muestra las tasas de desempleo del primer trimestre de 2003 en un conjunto de países. Los problemas 5, 6 y 7 se basan en ella.

Estados Unidos	Australia	Canadá	Alemania	Italia	Japón	Suecia
6,2	6,1	6,9	9,2	8,9	5,4	6,1

5. Supongamos que se ha seleccionado una muestra aleatoria de 400 trabajadores alemanes. Aproxime la probabilidad de que:
   (a) Como máximo, 40 de ellos estén desempleados.
   (b) Más de 50 estén desempleados.
6. Supongamos que se ha seleccionado una muestra aleatoria de 600 trabajadores japoneses. Aproxime la probabilidad de que:
   (a) Como máximo, 30 de ellos no tengan empleo.
   (b) Más de 40 estén desempleados.
7. Supongamos que se ha seleccionado una muestra aleatoria de 200 trabajadores canadienses. Aproxime la probabilidad de que:
   (a) Como máximo, 10 de ellos estén desempleados.
   (b) Más de 25 no tengan empleo.
8. Si el 65% de la población de una determinada comunidad está a favor de una propuesta de aumento de las tasas escolares, calcule la probabilidad de que, en una muestra aleatoria de 100 individuos de la comunidad:
   (a) Al menos 45 que estén a favor de la propuesta.
   (b) Haya menos de 60 que estén a favor.
   (c) Haya entre 55 y 75 que estén a favor.
9. El tamaño ideal de una clase de primer año en una determinada universidad es de 160 estudiantes. Por su experiencia, la universidad sabe que solamente un 40% de los admitidos asiste a clase. Basándose en esto, la universidad mantiene la política de aceptar 350 solicitudes. Mediante la aproximación normal, aproxime la probabilidad de que esto ocasione que:
   (a) Haya más de 160 estudiantes que asistan a clase.
   (b) Haya menos de 150 estudiantes que asistan a clase.
10. En una compañía aérea, el porcentaje de pasajeros que tienen reserva y no se presentan es del 6%. Si existen 260 personas con reserva en un determinado vuelo que puede admitir un máximo de 250 pasajeros, aproxime la probabilidad de que la compañía sea capaz de acomodar a todos los pasajeros con reserva que aparezcan.

La tabla siguiente muestra la lista de áreas de estudio y los porcentajes de alumnos que eligen cada una de ellas en una determinada universidad. Los problemas del 11 al 14 se basan en esta tabla y se supondrá, en todos estos problemas, que se ha seleccionado una muestra aleatoria de 200 alumnos que acaban de comenzar sus estudios en dicha universidad.

Área de estudio	**Porcentaje**
Artes y Humanidades	9
Ciencias Biológicas	4
Economía	27
Educación	9
Ingeniería	10
Ciencias Físicas	2
Ciencias Sociales	9
Estudios Profesionales	11
Estudios Técnicos	3
Otros estudios	16

Fuente: Instituto de Educación Superior. Universidad de California, Los Ángeles, *Anuario*.

11. ¿Cuál es la probabilidad de que 22 o más estudiantes de la muestra estudien Artes y Humanidades?

12. ¿Cuál es la probabilidad de que más de 60 de los estudiantes seleccionados estudien Economía?

13. ¿Cuál es la probabilidad de que 30 o más de los estudiantes de la muestra estudien una de las Ciencias (Biológicas, Físicas o Sociales)?

14. ¿Cuál es la probabilidad de que menos de 15 los estudiantes de la muestra estudien una Ingeniería?

15. Sea X una variable aleatoria binomial con parámetros $n = 100$ y $p = 0{,}2$. Calcule las probabilidades siguientes:

    (a) $P\{X \leq 25\}$

    (b) $P\{X > 30\}$

    (c) $P\{15 < X < 22\}$

16. Sea X una variable aleatoria binomial con parámetros $n = 150$ y $p = 0{,}6$. Calcule las probabilidades siguientes:

    (a) $P\{X \leq 100\}$

    (b) $P\{X > 75\}$

    (c) $P\{80 < X < 100\}$

17. En un estudio reciente se prueba que el 54% de los estudiantes que entran en la universidad no finalizan sus estudios universitarios en los cuatro años previstos. Supongamos que se ha seleccionado una muestra aleatoria de 500 alumnos que acaban de iniciar sus estudios universitarios.

    (a) ¿Cuál es la probabilidad aproximada de que menos de la mitad de ellos hayan terminado al cabo de 4 años?

    (b) ¿Cuál es la probabilidad aproximada de que más de 175 y menos de 225 hayan terminado al cabo de 4 años?

La tabla siguiente muestra los porcentajes de individuos, por sexo, que tienen ciertos hábitos de conducta perjudiciales para la salud. Los problemas 18, 19 y 20 se refieren a ella.

	Duermen 6 o menos horas al día	**Fuman**	**Nunca desayunan**	**Tienen un sobrepeso superior al 30%**
Hombres	22,7	32,6	25,2	12,1
Mujeres	21,4	27,8	23,6	13,7

Fuente: Centro Nacional de Estados Unidos de Estadísticas de la Salud, *Promoción de la Salud y Prevención de las Enfermedades*. 1985.

18. Supongamos que se elige una muestra aleatoria de 300 hombres. Aproxime la probabilidad de que:

    (a) Al menos 75 nunca desayunen.

    (b) Fumen menos de 100.

19. Supongamos que se extrae una muestra aleatoria de 300 mujeres. Aproxime la probabilidad de que:

    (a) Al menos 25 de ellas tengan un sobrepeso superior al 30%.

    (b) Menos de 50 duerman 6 horas como máximo.

20. Supongamos que se extraen dos muestras de 300 hombres y de 300 mujeres, respectivamente. Calcule la probabilidad aproximada de que haya más fumadores en la muestra de hombres que en la de mujeres. (*Sugerencia*: Denote por X e Y, respectivamente, el número de hombres y de mujeres que fuman en cada una de las muestras. Escriba la probabilidad pedida como $P\{X-Y>0\}$, y recuerde que la diferencia de dos variables aleatorias normales independientes es también una variable aleatoria normal.)

7.6 Distribución de la varianza muestral en una población normal

Antes de determinar la distribución de la varianza muestral cuando se muestrea sobre una población normal es necesario introducir la distribución chi-cuadrado, que se define como la distribución de la suma de los cuadrados de varias variables aleatorias normales estándar e independientes.

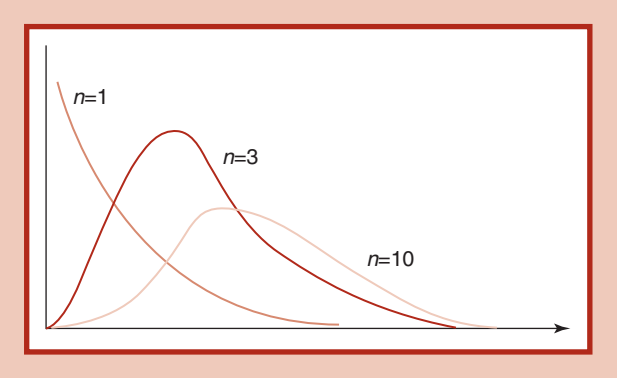


Figura 7.5 Funciones de densidad de la chi-cuadrado con n grados de libertad, para $n = 1, 3, 10$.

Definición

Si $Z_1, \ldots, Z_n$ son variables aleatorias normales estándar e independientes, la variable aleatoria

$$\sum_{i=1}^{n} Z_i^2$$

se dice que es una variable aleatoria *chi-cuadrado* con *n grados de libertad.*

La figura 7.5 representa las funciones de densidad de la chi-cuadrado para tres valores distintos de los grados de libertad n.

Para obtener la esperanza de una variable aleatoria chi-cuadrado, observe en primer lugar que, para una variable aleatoria, Z, normal estándar,

$$\begin{aligned} 1 &= Var(Z) \\ &= E[Z^2] - (E[Z])^2 \\ &= E[Z^2] \qquad \text{puesto que } E[Z] = 0 \end{aligned}$$

De aquí se desprende que $E[Z^2] = 1$ y, en consecuencia,

$$E\left[\sum_{i=1}^{n} Z_i^2\right] = \sum_{i=1}^{n} E[Z_i^2] = n$$

El valor esperado de una variable aleatoria chi-cuadrado es igual a su número de grados de libertad.

Supongamos ahora que se tiene una muestra $X_1, \ldots, X_n$ procedente de una población normal con media μ y varianza σ^2. Consideremos la varianza muestral S^2 definida por

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n - 1}$$

Se puede demostrar el teorema siguiente:

Teorema

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2}$$

sigue una distribución chi-cuadrado con $n - 1$ grados de libertad.

Pese a que la demostración matemática de este teorema excede el nivel de este libro, se puede llegar a entender por qué se cumple. Este razonamiento será de utilidad para guiar nuestra intuición a medida que se progrese en la lectura de los siguientes capítulos. Para empezar: consideremos las variables estandarizadas $(X_i - \mu)/\sigma$, $i = 1, \ldots, n$, donde μ representa la media poblacional. Puesto que todas estas variables son independientes y normales estándar, se tiene que la suma de sus cuadrados.

$$\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\sigma^2}$$

sigue una chi-cuadrado con n grados de libertad. Ahora bien, si se sustituye la media poblacional μ por la media muestral $\overline{X}$, la nueva variable aleatoria

$$\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2}$$

continúa siendo una chi-cuadrado, pero pierde un grado de libertad debido a que la media poblacional (μ) se ha reemplazado por su estimador (la media muestral $\overline{X}$).

Problemas

1. Los siguientes conjuntos de datos proceden de poblaciones normales, cuya desviación típica σ se especifica en cada caso. Obtenga el valor un estadístico cuya distri-

bución sea una chi-cuadrado, e indique el número de grados de libertad que tiene su distribución.

(a) 104, 110, 100, 98, 106; $\sigma = 4$

(b) 1,2, 1,6, 2,0, 1,5, 1,3, 1,8; $\sigma = 0{,}5$

(c) 12,4, 14,0, 16,0; $\sigma = 2{,}4$

2. Explique por qué una variable aleatoria chi-cuadrado con n grados de libertad tiene una distribución que se aproxima a la de una variable aleatoria normal, cuando n es suficientemente grande. (*Sugerencia*: Utilice el teorema central del límite.)

Términos clave

Muestra procedente de una distribución poblacional: Si $X_1, \ldots, X_n$ son variables aleatorias independientes con idéntica distribución F, se dice que forman una muestra procedente de la distribución F.

Estadístico: Variable numérica cuyo valor se puede determinar a partir de la muestra.

Media muestral: Si $X_1, \ldots, X_n$ es una muestra, la media muestral es

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Varianza muestral: Si $X_1, \ldots, X_n$ es una muestra, la varianza muestral es

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n - 1}$$

Teorema central del límite: Teorema que establece que la suma de una muestra de tamaño n procedente de una población sigue aproximadamente una distribución normal, si n es grande.

Muestra aleatoria: Una muestra de n miembros de una población es una muestra aleatoria si se extrae de tal forma que todos los posibles subconjuntos de n miembros de la población tienen la misma probabilidad de constituir la muestra.

Distribución chi-cuadrado con *n* grados de libertad: Distribución que sigue la suma de los cuadrados de n variables aleatorias normales estándar e independientes.

Resumen

Si $\overline{X}$ es la media muestral de una muestra de tamaño n procedente de una población con media μ y desviación típica σ, la media y desviación típica de $\overline{X}$ son

$$E[\overline{X}] = \mu \quad \text{y} \quad \text{SD}(\overline{X}) = \frac{\sigma}{\sqrt{n}}$$

El teorema central del límite establece que la media muestral de una muestra de tamaño n extraída de una población con media μ y desviación típica σ se distribuye aproximadamente, si n es grande, como una normal de media μ y desviación típica $\sigma/\sqrt{n}$.

Consideremos una muestra de tamaño n procedente de una población con N individuos, de los cuales Np presentan una determinada característica. Denotemos por X el número de elementos de la muestra que presentan la característica. Si N es grande comparado con n, X se distribuye aproximadamente como una variable aleatoria binomial de parámetros n y p. De aquí en adelante, en lo que resta del libro, se asumirá que X sigue esta última distribución.

Se desprende de lo anterior que la proporción de individuos de la muestra que presentan la característica, es decir, $\overline{X} = X/n$, tendrá una media y una desviación típica dadas por

$$E[\overline{X}] = p \quad \text{y} \quad \text{SD}(\overline{X}) = \sqrt{\frac{p(1-p)}{n}}$$

Del teorema central del límite se desprende que una variable aleatoria binomial con parámetros n y p sigue aproximadamente, si n es grande, una distribución normal con media np y desviación típica $\sqrt{np(1-p)}$. Esta aproximación es buena si n es lo suficientemente grande como para que np y $n(1-p)$ sean mayores que 5.

Si S^2 es la varianza muestral de una muestra de tamaño n procedente de una población normal con varianza σ^2, se verifica que $(n-1)S^2/\sigma^2$ sigue una distribución chi-cuadrado con $n-1$ grados de libertad.

El valor esperado de una variable aleatoria chi-cuadrado es igual al número de sus grados de libertad.

Problemas de repaso

1. La media muestral y la desviación típica muestral de las puntuaciones de todos los estudiantes que se han presentado a un determinado test fueron, respectivamente, 517 y 120. Si se extrae una muestra aleatoria de 144 de los citados estudiantes, calcule la probabilidad aproximada de que la puntuación media de la muestra sobrepase:

   (a) 507

   (b) 517

   (c) 537

   (d) 550

2. Sea $\overline{X}$ la media muestral de una muestra de tamaño 10 procedente de una población cuya distribución de probabilidad viene dada por

$$P\{X = i\} = \begin{cases} 0{,}1 & \text{si } i = 1 \\ 0{,}2 & \text{si } i = 2 \\ 0{,}3 & \text{si } i = 3 \\ 0{,}4 & \text{si } i = 4 \end{cases}$$

Obtenga

(a) La media poblacional μ

(b) La desviación típica poblacional σ

(c) $E[\overline{X}]$

(d) $\text{Var}(\overline{X})$

(e) $\text{SD}(\overline{X})$

3. En el problema 2, supongamos que el tamaño muestral es 2. Calcule la distribución de probabilidad de $\overline{X}$ y utilícela después para calcular $E(\overline{X})$ y $\text{SD}(\overline{X})$. Compruebe las respuestas a partir de los valores de μ y σ.

4. La media y la desviación típica de las duraciones de las baterías utilizadas en determinados coches eléctricos son, respectivamente, 225 y 24 minutos. Aproxime la probabilidad de que la duración total de un conjunto de 10 baterías, usadas una tras otra, sea superior a:

   (a) 2200 minutos

   (b) 2350 minutos

   (c) 2500 minutos

   (d) ¿Cuál es la probabilidad de que la duración total esté comprendida entre 2200 y 2350 minutos?

5. Supongamos que el 12% de los miembros de una población son zurdos. Sobre una muestra aleatoria de 100 individuos de la población

   (a) Calcule la media y la desviación típica del número de zurdos.

   (b) Obtenga la probabilidad de que ese número esté comprendido entre 10 y 14, ambos inclusive.

6. El peso de una persona elegida aleatoriamente entre los pasajeros de un trasbordador tiene por media 155 libras, con una desviación típica de 28 libras. Si la capacidad del trasbordador es de 100 pasajeros y éste va completo, calcule la probabilidad de que la carga del pasaje supere 16 000 libras.

7. Las facturas mensuales de teléfono de los estudiantes de una determinada residencia universitaria tienen un valor esperado de 15 dólares y una desviación típica de 7 dólares. Denotemos por X la suma de las facturas telefónicas mensuales de una muestra de 20 de los estudiantes citados.

   (a) ¿Cuál es el valor de $E(X)$?

   (b) ¿Cuál es el valor de $\text{SD}(X)$?

   (c) Calcule la probabilidad de que X sobrepase 300 dólares.

8. En un artículo de periódico se mantiene que el salario medio de los ingenieros químicos recién graduados es de 54 000 dólares, con una desviación típica de 5000 dólares. Supongamos que se selecciona una muestra aleatoria de 12 de dichos ingenieros y que

su salario medio resultó ser 45 000 dólares. Si se asume que las cifras que cita el artículo son correctas ¿cuál es la probabilidad de que con una muestra como la anterior se observe un salario medio muestral inferior o igual a 45 000 dólares?

9. Una agencia de publicidad ha puesto en marcha una campaña de promoción de un nuevo producto. Al final de la misma, la agencia mantiene que al menos un 25% del total de los consumidores conoce el producto. Para comprobar esto, el productor selecciona aleatoriamente a 1000 consumidores y observa que 232 de ellos conocen el producto. Si realmente el 25% de los consumidores conocieran el producto, ¿cuál es la probabilidad de que, en una muestra de 1000 consumidores, existan como máximo 232 de ellos que conocen el producto?

10. Un equipo de baloncesto de un club juega 60 partidos por temporada. De esos partidos, 32 se juegan contra equipos de nivel A y 28 contra equipos de nivel B. Asumamos que los resultados de los distintos partidos son independientes y que el equipo en cuestión gana con unas probabilidades de 0,5 y 0,7 a sus oponentes de nivel A y B, respectivamente. Denotemos como X al número total de partidos ganados por temporada.

    (a) ¿Es X una variable aleatoria binomial?

    (b) Denotemos por X_A y X_B a los partidos ganados por el equipo contra oponentes de nivel A y de nivel B, respectivamente. ¿Cómo se distribuyen X_A y X_B?

    (c) ¿Qué relación existe entre X_A, X_B y X?

    (d) Calcule la probabilidad de que el equipo gane al menos 40 partidos en una temporada. (*Sugerencia*: Recuerde que la suma de de dos variables aleatorias normales e independientes es también una variable aleatoria normal.)

11. Si X es una binomial de parámetros $n = 80$ y $p = 0{,}4$, calcule las probabilidades siguientes:

    (a) $P\{X > 34\}$

    (b) $P\{X \leq 42\}$

    (c) $P\{25 \leq X \leq 39\}$

12. Consideremos el siguiente modelo simple referido a los cambios diarios en el precio de una acción. Supongamos que cada día el precio de la acción puede subir una unidad con una probabilidad de 0,52 o bajar una unidad con una probabilidad de 0,48. Si el precio de la acción en el día de hoy es 200 y se denota por X el precio que tendrá al cabo de 100 días,

    (a) Defina las variables aleatorias $X_1, X_2, \ldots, X_{100}$ tales que

$$X = 200 + \sum_{i=1}^{100} X_i$$

    (b) Obtenga $E[X_i]$.

    (c) Calcule $\text{Var}(X_i)$.

    (d) Utilice el teorema central del límite para calcular $P\{X \geq 210\}$.

13. Se muestran a continuación los porcentajes de residentes en Estados Unidos que no disponían de seguro médico en 2002, clasificados por edad.

AEdad	Porcentaje sin seguro
Menos de18	11,6
De 18 a 24	29,6
De 25 a 34	24,9
De 35 a 44	17,7
De 45 a 64	13,5
De 65 o más	0,8

Supongamos que se extrae una muestra aleatoria de tamaño 1000 de cada una de las clases de edad. Calcule aproximadamente la probabilidad de que:

(a) Al menos 100 de los individuos de la muestra menores de 18 años no tengan seguro médico.

(b) Haya menos de 260 personas sin seguro en la muestra entre los individuos cuya edad está comprendida entre 25 y 34 años.

(c) No tengan seguro como mínimo 5 entre los mayores de 64 años, y como máximo 120 entre aquellos cuya edad está entre 45 a 64 años.

(d) Haya más personas sin seguro en la muestra con edades comprendidas entre 18 y 24 años que en la muestra con edades de 25 a 34 años.

14. Un administrador de una universidad pretende estimar con rapidez el número medio de estudiantes por clase. Debido a que no quiere que los profesores conozcan su intención, decide solicitar la ayuda de los estudiantes. Para ello ha extraído aleatoriamente 100 nombres del registro de alumnos de la universidad, y les pide que averigüen y le informen del número de estudiantes de sus clases. Después, el administrador decide estimar el número medio de estudiantes por clase como la media de los 100 valores dados por los estudiantes.

(a) ¿Consigue con este método el objetivo deseado?

(b) Si la contestación al apartado (a) es afirmativa, explique por qué; si es negativa, proponga un procedimiento que funcione mejor.

8

Estimación

¡Datos, datos, datos! –gritó impacientemente–. No puedo hacer ladrillos sin arcilla.

Sherlock Holmes (A. C. Doyle), *Las aventuras de los bombachos de cobre*

Se verá cómo se pueden utilizar los datos muestrales para estimar una media poblacional, una varianza poblacional y una proporción poblacional. Se analizarán los estimadores puntuales, que son estimadores que asignan un solo valor al parámetro. Se tendrá en cuenta el error estándar de esos estimadores. También se estudiarán los estimadores por intervalo, que contienen el parámetro con un nivel de confianza dado.

8.1 Introducción

No es raro leer en un periódico frases como: "Una encuesta reciente de 1500 americanos elegidos aleatoriamente pone de manifiesto que el 22 por ciento de la población total de Estados Unidos está a dieta, con un margen de error de ± 2 por ciento." Quizá el lector se habrá preguntado acerca del significado de estos términos. Por ejemplo, ¿qué significa exactamente *con un margen de error de* ± 2 *por ciento*? Y también, ¿cómo es posible que, a partir de una muestra de solamente 1500 adultos, se pueda obtener la proporción de personas adultas que están a dieta en un país con más de 150 millones de adultos?

En este capítulo se encontrarán las respuestas a estas cuestiones. En general, se verá cómo se puede obtener información acerca de una característica numérica de una población mediante el análisis de los resultados obtenidos a partir de una muestra extraída de dicha población.

Aunque se pueden sintetizar los valores numéricos de los miembros de una población mediante una distribución de probabilidad poblacional, esta distribución no suele conocerse al completo. Por ejemplo, algunos de sus parámetros, tales como su media o su desviación típica, pueden ser desconocidos. Un punto de interés fundamental en la Estadística consiste en cómo se pueden utilizar los resultados de una muestra extraída de la población para estimar dichos parámetros poblacionales desconocidos.

Por ejemplo, si los elementos de la población consisten en los chips de ordenador producidos recientemente, uno podría estar interesado en saber cuál es la vida media de funcionamiento de dichos chips. Es decir, uno podría querer estimar la media poblacional de la distribución de las duraciones de los chips.

En este capítulo se presentarán distintas formas de estimar determinados parámetros de la distribución poblacional. Para ello, se verá cómo se pueden utilizar los estimadores y los valores estimados obtenidos a partir de aquéllos.

Definición

Un *estimador* es un estadístico cuyos valores dependen de la muestra particular extraída. Se utiliza el valor del estimador, llamado *valor estimado*, para predecir el valor de un parámetro poblacional de interés.

Por ejemplo, si se desea estimar la vida media de los chips, se podría emplear la media muestral como *estimador* de la media poblacional. Si el valor de la media muestral fuera de 122 horas, el valor estimado de la media poblacional sería 122 horas.

En la sección 8.2 se considerará el problema de cómo estimar una media poblacional, y en la sección 8.3 se analizará cómo estimar una proporción poblacional. La sección 8.4 aborda el problema de cómo estimar una varianza poblacional. Los estimadores considerados en estas secciones se denominan *estimadores puntuales*, porque proporcionan un valor único que interesa que esté próximo al parámetro que se desea estimar. En las restantes secciones, se estudiará el problema de obtener *estimadores por intervalo*. En este caso, en lugar de proporcionar un solo valor como estimador, se intenta obtener un intervalo que contenga el parámetro en cuestión. También se analizará cómo se puede asignar una determinada confianza a un estimador por intervalo dado; es decir, cómo se puede asignar una determinada certidumbre al hecho de que el parámetro caiga realmente dentro del intervalo estimado.

8.2 Estimador puntual de la media de una población

Denotemos por $X_1, \ldots, X_n$ a una muestra extraída de una población con una media desconocida μ. Se puede utilizar la media muestral $\overline{X}$ como estimador de μ, puesto que, como se vio en la sección 7.3,

$$E[\overline{X}] = \mu$$

se tiene que el valor esperado de este estimador coincide con el valor que se desea estimar. Los estimadores que cumplen esta condición se denominan *insesgados*.

Definición

Un estimador cuyo valor esperado coincide con el parámetro que se desea estimar se dice que es un estimador *insesgado* de dicho parámetro.

Ejemplo 8.1 Para estimar la cantidad media reclamada por incendios en apartamentos de tamaño medio, una organización de consumidores muestreó los ficheros de una gran compañía de seguros y obtuvo las siguientes cantidades (en miles de dólares) que habían sido reclamadas en 10 incendios:

$$121, 55, 63, 12, 8, 141, 42, 51, 66, 103$$

El estimador de la media de las cantidades reclamadas por daños sobre el total de incendios es, pues,

$$\begin{aligned}\overline{X} &= \frac{121 + 55 + 63 + 12 + 8 + 141 + 42 + 51 + 66 + 103}{10} \\ &= \frac{662}{10} = 66{,}2\end{aligned}$$

Es decir, se estima que la cantidad media reclamada por incendio es de 66 200 $. ■

Como se ha visto, el valor esperado de la media muestral $\overline{X}$ es μ. Puesto que no es muy probable que una variable aleatoria difiera mucho de la media poblacional en unidades de su desviación típica, es importante tener en cuenta la desviación típica de $\overline{X}$. En la sección 7.3 se vio que

$$\mathrm{SD}(\overline{X}) = \frac{\sigma}{\sqrt{n}}$$

donde σ representa la desviación típica de la población. El valor de $\mathrm{SD}(\overline{X})$ habitualmente se denomina *error estándar* del estimador de la media poblacional $\overline{X}$. Puesto que es muy improbable que una variable aleatoria se separe de su media más de 2 veces su desviación típica (máxime cuando la variable aleatoria es aproximadamente normal, tal como ocurre con $\overline{X}$ cuando el tamaño muestral es suficientemente grande), se puede tener una gran confianza en que el estimador de la media poblacional difiera del valor correcto en menos ± 2 veces su error estándar. Observe que el error estándar viene dado por un cociente en cuyo denominador figura la raíz cuadrada del tamaño muestral; en consecuencia, para reducir a la mitad el error estándar se debe multiplicar por 4 el tamaño de la muestra.

Ejemplo 8.2 Para un mismo individuo se han de realizar distintas mediciones de su nivel de potasio en la sangre debido tanto a la imprecisión del procedimiento de medición como al hecho de que el nivel real varía, dependiendo de factores tales como la cantidad de ali-

mentos consumidos o los ejercicios físicos realizados recientemente. Supongamos que, para un individuo concreto se sabe que sus niveles de potasio varían alrededor de un valor medio μ con una desviación típica de 0,3. Si en cuatro mediciones de su nivel de potasio se han obtenido los valores siguientes

$$3{,}6,\ 3{,}9,\ 3{,}4,\ 3{,}5$$

se tiene que el estimador del nivel medio de potasio en la sangre para la persona en cuestión es

$$\frac{3{,}6 + 3{,}9 + 3{,}4 + 3{,}5}{4} = 3{,}6$$

y el error estándar de este estimador es igual a

$$\mathrm{SD}(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \frac{0{,}3}{2} = 0{,}15$$

Por consiguiente, se puede tener una confianza elevada en que el valor real de la media no difiera de 3,6 en más de 0,30.

Supongamos que se quiere que el estimador tenga un error estándar de 0,05. Esto supone dividir el error estándar por 3, y por ello se debería extraer una muestra con un tamaño 9 veces mayor. Esto es, se deberían realizar 36 mediciones de los niveles de potasio del individuo en cuestión. ■

Problemas

1. Los pesos de los miembros de una muestra aleatoria de ocho participantes en el maratón de Boston de 2004 fueron los siguientes:

   $$121,\ 163,\ 144,\ 152,\ 186,\ 130,\ 128,\ 140$$

   Utilice estos datos para estimar el peso medio de todos los participantes en la carrera citada.

2. Supongamos que los datos del problema 1 representan los pesos de los ocho primeros clasificados en la carrera. ¿Sería posible utilizar estos datos para estimar el peso medio de todos los corredores? ¡Explique la respuesta!

3. Para determinar el gasto medio en libros de texto de los estudiantes de una determinada universidad se seleccionó una muestra de 10 estudiantes que fueron entrevistados posteriormente. Si las cantidades gastadas (redondeadas al entero más próximo) fueron

   $$422,\ 146,\ 368,\ 52,\ 212,\ 454,\ 366,\ 711,\ 227,\ 680$$

   ¿cuál es el estimador del gasto medio en libros de texto para el total de estudiantes de dicha universidad?

4. De una muestra de nueve alumnos de preescolar en una determinada zona residencial se obtuvieron los datos siguientes, relativos al número de horas al día que empleaba cada uno de ellos en ver la televisión,

   3, 0, 5, 3,5, 1,5, 2, 3, 2,5, 2

   Para la población de alumnos de preescolar del citado barrio residencial, estime el número medio de horas por día que emplean en ver la televisión.

5. Una compañía que fabrica reproductores de discos compactos desea estimar la vida media de los lectores de láser de sus aparatos. Por este motivo se selecciona una muestra de 40 lectores. Si la suma de las duraciones de esos lectores de láser fue de 6624 horas, ¿cuál es el valor estimado de la vida media de todos los lectores de láser de esos aparatos?

6. Para estimar el nivel medio de colesterol en sangre para la población de trabajadores adultos se ha extraído una muestra de 1000 trabajadores. Si se desea reducir el error estándar a una cuarta parte, ¿qué tamaño muestral se precisa?

7. Se sabe que la desviación típica de los pesos de los recién nacidos es de 10 onzas. Si se desea estimar el peso medio de todos los recién nacidos, ¿qué tamaño muestral es necesario para que el error estándar del estimador sea menor que 3 onzas?

8. Los siguientes datos muestran los tiempos que tuvieron que esperar los miembros de una muestra aleatoria de 12 pacientes en una determinada consulta médica de un hospital:

   46, 38, 22, 54, 60, 36, 44, 50, 35, 66, 48, 30

   Utilice estos datos para estimar el tiempo medio de espera para el total de pacientes de dicha consulta médica.

9. En la siguiente tabla de frecuencias se reflejan los tamaños de los hogares unifamiliares de una muestra de éstos en una ciudad determinada.

Tamaño del hogar	**Frecuencia**
1	11
2	19
3	28
4	26
5	11
6	4
7	1

   Estime el tamaño medio de todos los hogares unifamiliares de esa ciudad.

10. ¿En cuál de los apartados (a) o (b) se obtiene un estimador de μ más preciso?

    (a) La media muestral de una muestra de tamaño n extraída de una población con media μ y varianza σ^2.

(b) La media muestral de una muestra de tamaño $3n$ extraída de una población con media μ y varianza $2\sigma^2$.

(c) ¿Cuál tendría que ser el tamaño de la muestra en (b), para que el estimador allí obtenido iguale la precisión del estimador de (a)?

11. Repita el problema 10 si (a) y (b) son los siguientes:

(a) La media muestral de una muestra de tamaño n extraída de una población con media μ y desviación típica σ.

(b) La media muestral de una muestra de tamaño $3n$ extraída de una población con media μ y desviación típica 3σ.

8.3 Estimador puntual de una proporción poblacional

Supongamos que se intenta estimar la proporción de individuos de una población que está a favor de una determinada propuesta. Denotemos por p a dicha proporción desconocida. Para estimar p, en primer lugar, se debería seleccionar una muestra aleatoria y, después, utilizar como estimador de p la proporción de elementos de la muestra que están a favor de la propuesta. Si este estimador se representa por $\hat{p}$, se tendría que

$$\hat{p} = \frac{X}{n}$$

siendo X el número de individuos de la muestra que están a favor de la propuesta, y n el tamaño muestral.

A partir de los resultados de la sección 7.5 se sabe que

$$E[\hat{p}] = p$$

Esto es, la proporción, $\hat{p}$, de elementos de la muestra a favor de la propuesta es un estimador insesgado de p, la proporción de miembros de la población total a favor de la propuesta. La dispersión del estimador $\hat{p}$ alrededor de su media p se mide mediante su desviación típica, que (como se vio en la sección 7.5) es igual a

$$\mathrm{SD}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

La desviación típica de $\hat{p}$ se denomina también *error estándar* de $\hat{p}$ como estimador de la proporción poblacional p. Se desprende de la fórmula anterior que este error estándar será pequeño siempre que el tamaño muestral n sea grande. De hecho, se puede demostrar que, para cualquier valor de p, se verifica que

$$p(1-p) \leq \frac{1}{4}$$

de donde se sigue que

$$\mathrm{SD}(\hat{p}) \leq \sqrt{\frac{1}{4n}} = \frac{1}{2\sqrt{n}}$$

Por ejemplo, supongamos que se ha seleccionado una muestra aleatoria de tamaño 900. Independientemente de cuál sea el valor real de la proporción poblacional que está a favor de la propuesta, se tiene que el error estándar del estimador de esa proporción es menor o igual que $1/(2\sqrt{900}) = 1/60$.

En la expresión anterior, que acota el error estándar del estimador, se asume que la muestra de tamaño n se ha extraído de una población infinitamente grande. Cuando la población es finita (tal como, naturalmente, ocurre en la práctica), el error estándar es menor, con lo que el estimador será aún más preciso que lo indicado anteriormente.

Ejemplo 8.3 En un colegio se intenta determinar la reacción de los estudiantes a cierta norma sobre la forma de vestir. Para ello, el colegio ha seleccionado una muestra aleatoria de 50 estudiantes que, posteriormente, fueron entrevistados. Si 20 de estos estudiantes se manifestaron a favor de la propuesta:

(a) Estime la proporción estudiantes, sobre el total de éstos, que están a favor de la propuesta.

(b) Estime el error estándar de este estimador.

Solución

(a) El estimador de la proporción poblacional de estudiantes a favor de la norma en la forma de vestir es $20/50 = 0{,}40$.

(b) El error estándar de ese estimador es $\sqrt{p(1-p)/50}$, donde p representa la proporción poblacional de estudiantes a favor de la norma. Utilizando 0,4 como valor estimado de p se puede estimar el error estándar mediante $\sqrt{0{,}4(1-0{,}4)/50} = 0{,}0693$. ■

Problemas

1. En 1985, de una muestra aleatoria de 1325 norteamericanos, 510 opinaron que, si se convocaran unas elecciones libres en la Unión Soviética, el partido comunista las ganaría. Estime la proporción de personas sobre el total de norteamericanos que eran de esa misma opinión en dicha época.

2. Estime el error estándar del estimador obtenido en el problema 1.

3. Para conocer el porcentaje de socios de una gran organización social que están a favor de aumentar la cuota anual se entrevistó a una muestra aleatoria de 20 socios de la organización. Si 13 de esos socios se manifestaron a favor del aumento, ¿cuál es el estimador de la proporción de socios, sobre el total de éstos, que están a favor? ¿Cuál es el valor estimado del error estándar del estimador citado?

4. Se muestran a continuación los resultados de 20 juegos de un solitario de cartas, en los que sólo se puede ganar (w) o perder (l).

$$w, l, l, l, w, l, l, w, l, w, w, l, l, l, l, w, l, l, w, l$$

   (a) Estime la probabilidad de ganar cuando se juega a ese solitario.

   (b) Estime el error estándar del estimador obtenido en el apartado (a).

5. Sobre una muestra de 85 estudiantes de una gran universidad pública se observó que 35 de ellos tenían un coche con menos de 5 años de antigüedad. Estime la proporción de estudiantes de la universidad que tienen coches de dicha característica. ¿Cuál es el valor estimado del error estándar del estimador propuesto?
6. Sobre una muestra de 100 padres se observó que 64 de ellos estaban a favor de aumentar la edad para poder obtener el permiso de circulación.
   (a) Sobre la población total de padres estime la proporción que están a favor de elevar la edad de conducir.
   (b) Estime el error estándar del estimador.
7. En una muestra aleatoria de 1000 trabajadores de la construcción, 122 de ellos estaban desempleados.
   (a) Estime la proporción de trabajadores de la construcción que están en paro.
   (b) Estime el error estándar del estimador obtenido en el apartado (a).
8. En una muestra aleatoria de 500 arquitectos, 104 eran mujeres.
   (a) Estime la proporción de mujeres sobre el total de arquitectos.
   (b) Estime el error estándar del estimador obtenido en el apartado (a).

9. Una muestra aleatoria de 1200 ingenieros de Estados Unidos incluía a 28 de origen hispano, a 45 de origen africano y a 104 mujeres. Sobre el total de ingenieros, estime la proporción de ellos que son:
   (a) de origen hispano
   (b) de origen africano
   (c) mujeres
10. Para los apartados (a), (b) y (c) del problema 9, estime el error estándar del estimador.
11. En una muestra de 400 certificados de defunción de jóvenes menores de 20 años se observó que en 98 de ellos las muertes habían sido causadas por accidentes de tráfico.
   (a) Sobre el total de fallecimientos de jóvenes menores de 20 años, calcule la proporción de muertes cuyas causas de defunción sean los accidentes de tráfico.
   (b) Estime el error estándar del estimador obtenido en el apartado (a).

12. Se está diseñando una encuesta para conocer la proporción poblacional de individuos que están a favor de una determinada medida escolar. ¿Cuál debe ser el tamaño muestral para que se pueda asegurar que el error estándar del estimador resultante es menor o igual que 0,1?

13. La ciudad de Los Ángeles tiene aproximadamente 3 veces más votantes que San Diego. En cada una de estas ciudades se planea realizar una consulta popular acerca de una determinada medida escolar. Para anticipar la opinión de los votantes se han seleccionado sendas muestras de votantes en cada una de las ciudades, cuyos tamaños muestrales fueron 3000 para la muestra de Los Ángeles y 1000 para la de San Diego. De las siguientes sentencias, ¿cuál se cree que es más precisa?

    (a) Los estimadores de la proporción poblacional de personas a favor de la medida tienen la misma precisión en las dos ciudades.

    (b) El estimador de Los Ángeles tiene 3 veces más precisión que el de San Diego.

    (c) El estimador de Los Ángeles tiene aproximadamente 1,7 veces más precisión que el de San Diego.

    Explique cómo se interpreta la palabra *precisión* en las sentencias (a), (b) y (c).

*14. En la ciudad de Chicago existían 12 048 policías a jornada completa en 1990. Para averiguar cuántos afroamericanos existían en el grupo se seleccionó una muestra aleatoria de 600 policías y se observó que 87 de ellos eran de origen africano.

    (a) Estime el número de afroamericanos que había en el grupo total de policías.

    (b) Estime el error estándar del estimador obtenido en el apartado (a).

*8.3.1 Estimación de la probabilidad de sucesos íntimos

Supongamos que a una compañía la interesa saber en qué medida sus empleados consumen drogas ilegales. La compañía reconoce que no es muy factible pensar que sus empleados contesten con veracidad a preguntas sobre este tema, incluso aunque se les haya garantizado que sus contestaciones serán secretas y no tendrán efecto alguno. La realidad es que, pese a que se asegure que las respuestas serán secretas y que no será posible identificar a quién las emite, los empleados pueden sospechar lo contrario y no contestar verazmente. Sobre esta base, ¿cómo puede la compañía conseguir la información que desea?

A continuación se presentará un método que permitirá obtener la información deseada y que, al mismo tiempo, protegerá la privacidad de los encuestados. El método se basa en una técnica de aleatorización que funciona como se describe a continuación. Para empezar, supongamos que la pregunta íntima se plantea de tal forma que la contestación *sí* pueda ser conflictiva. Por ejemplo, la pregunta se podría plantear: ¿Ha consumido drogas ilegales en el último mes? Presumiblemente, si la respuesta verdadera es *no*, el trabajador no dudará en darla. Si, por el contrario, la respuesta verdadera fuera *sí*, algunos trabajadores contestarían *no*, falseando su respuesta. Para evitar la propensión a mentir se puede proceder como se indica a continuación, tras habérselo explicado a los trabajadores. Una vez que se ha planteado la pregunta, el trabajador lanza una moneda sin que el encuestador observe el resultado obtenido. Si sale cara, el trabajador debe contestar *sí* a la pregunta íntima, con

independencia de cuál sea la realidad; mientras que si sale cruz, el trabajador debe contestar honestamente. Se ha de explicar al trabajador que una contestación *sí* no significa que se esté admitiendo haber consumido drogas ilegales, puesto que esta respuesta se daría si en el lanzamiento de la moneda se hubiera obtenido cara (lo cual ocurre en un 50% de los casos). De esta forma, se asegura que los trabajadores encuestados pueden contestar verazmente sin que su privacidad se vea amenazada.

Analicemos esta situación para ver cómo se puede estimar p, la proporción de trabajadores que han consumido drogas ilegales en el último mes. Sea $q = 1 - p$ la proporción de trabajadores que no han consumido drogas. Empezaremos calculando la proporción poblacional de trabajadores que contestarían *no*. Dado que esta respuesta se daría si (1) en el lanzamiento de la moneda resultó cruz y (2) el trabajador no consumió drogas ilegales en el último mes, se ve que

$$P\{\text{no}\} = \frac{1}{2} \times q = \frac{q}{2}$$

Por consiguiente, se puede tomar la proporción muestral de respuestas *no* como un estimador de $q/2$; o, equivalentemente, se puede estimar q mediante el doble de la proporción muestral de respuestas *no*. Puesto que $p = 1 - q$, con la anterior información muestral también se puede estimar p, la proporción poblacional de trabajadores que han consumido drogas ilegales en el pasado mes.

Por ejemplo, si el 70% de los trabajadores muestreados contestaron afirmativamente a la pregunta íntima, el 30% contestaron *no*; así pues, se estimará q mediante $2(0{,}3) = 0{,}6$. Esto es, se estimaría que el 60% de la población no ha consumido drogas ilegales durante el mes anterior y, en consecuencia, el 40% sí que lo habría hecho. Si el 35% de los trabajadores encuestados hubiera contestado *no*, se habría estimado que q es igual a $2(0{,}35) = 0{,}70$ y, por tanto, que $p = 0{,}3$. De igual forma, si el 48% de los encuestados hubiera contestado *no*, nuestro estimador de p sería $1 - 2(0{,}48) = 0{,}04$.

En consecuencia, si cada individuo muestral lanza una moneda se puede obtener un estimador veraz de p. Sin embargo, el "precio" que se debe pagar es un incremento en el error estándar. De hecho, se puede demostrar que el error estándar del estimador de p es ahora $\sqrt{(1 + p)(1 - p)/n}$, que resulta ser mucho mayor que el error estándar del estimador propuesto cuando no es necesario llevar a cabo los lanzamientos de la moneda (es decir, cuando se contesta honestamente, pues no se trata de una pregunta íntima).

Problemas

1. Supongamos que se está empleando el esquema de aleatorización descrito en este apartado. Si en una muestra de 50 personas resultaron 32 respuestas *sí*, ¿cuál sería el estimador de p?

2. En el problema 1, si 40 de las 50 personas hubieran contestado *sí*, ¿cuál sería el estimador de p?

3. Cuando se utiliza la técnica de aleatorización, el error estándar del estimador de p es $\sqrt{(1 + p)(1 - p)/n}$. Ahora bien, si la aleatorización no fuera necesaria porque todos

los encuestados contestaron verazmente, el error estándar del estimador de p sería $\sqrt{p(1-p)/n}$. La razón entre estos dos errores estándar es, pues,

$$\frac{\text{Error estándar con aleatorización}}{\text{Error estándar habitual}} = \sqrt{\frac{1+p}{p}}$$

Este cociente es, por tanto, un indicador del precio que se debe pagar si se trata de una pregunta íntima.

(a) ¿Es este precio mayor para valores grandes o para valores pequeños de p?

(a) Calcule el valor de este cociente para $p = 0{,}1$, $0{,}5$ y $0{,}9$.

8.4 Estimación de la varianza de una población

Supongamos que se ha extraído una muestra, $X_1, \ldots, X_n$, de tamaño n procedente de una población con varianza desconocida σ^2, y que se van a utilizar los datos muestrales para estimar σ^2. La varianza muestral S^2, definida por

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

es un estimador de la varianza poblacional σ^2. Para entender por qué es así, recordemos que la varianza de la población es la diferencia cuadrática esperada entre una observación y la media poblacional μ. Esto es, para $i = 1, \ldots, n$,

$$\sigma^2 = E[(X_i - \mu)^2]$$

Así pues, parece natural proponer como estimador de σ^2 la media de las diferencias al cuadrado entre los datos muestrales y la media de la población. Esto es, parece natural que un estimador apropiado de σ^2 sea

$$\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{n}$$

Realmente, es cierto que éste es un estimador apropiado de σ^2 cuando se conoce la media poblacional μ. Sin embargo, cuando la media μ de la población es desconocida también parece razonable utilizar la misma expresión anterior tras sustituir μ por su estimador $\overline{X}$. Para conseguir que, tras la sustitución, el nuevo estimador continúe siendo insesgado es preciso modificar el denominador cambiando n por $n-1$; de esta forma se obtiene el estimador S^2.

Si la media de la población μ es conocida, el estimador apropiado de la varianza poblacional σ^2 es

$$\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{n}$$

Cuando, por el contrario, la media de la población μ es desconocida, el estimador apropiado de la varianza poblacional σ^2 es

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n - 1}$$

S^2 es un estimador insesgado de σ^2, esto es,

$$E[S^2] = \sigma^2$$

Puesto que la varianza muestral S^2 se puede utilizar para estimar la varianza poblacional σ^2, es natural utilizar $\sqrt{S^2}$ para estimar la desviación típica poblacional σ.

La desviación típica poblacional σ se estima por S, la desviación típica muestral.

Ejemplo 8.4 En una muestra de nueve componentes electrónicos producidos por una determinada compañía se midió el tamaño (en unidades adecuadas) de cada componente, y como resultado se obtuvieron los datos siguientes:

1211, 1224, 1197, 1208, 1220, 1216, 1213, 1198, 1197

¿Cómo se estima la desviación típica poblacional y la varianza poblacional de los tamaños de dichos componentes?

Solución Para responder se ha de calcular la varianza muestral S^2. Puesto que restar una constante de cada dato muestral no afecta al valor de este estadístico empezaremos restando 1200 de cada dato para obtener el siguiente conjunto de datos transformados:

11, 24, 23, 8, 20, 16, 13, 22, 23

Si se utiliza un ordenador se obtiene que la varianza muestral y la desviación típica muestral de los datos transformados es

$$S^2 = 103 \qquad S = 10{,}149$$

En consecuencia, los valores estimados de la desviación típica poblacional y de la varianza poblacional son, respectivamente, 10,149 y 103. ■

La Estadística en perspectiva

La reducción de la varianza es la clave del éxito en la producción

De acuerdo con los expertos japoneses en control de calidad, la clave para que un proceso de producción tenga éxito –tanto si se están produciendo componentes de automóviles, equipos electrónicos, chips de ordenador, tornillos o cualquier otra cosa– consiste en asegurar que el proceso de producción obtiene con regularidad, a un precio razonable, productos cuyas características están próximas a las previamente fijadas como *objetivo*. Por ello entienden que, para cada unidad producida, existe un valor objetivo que el productor pretende conseguir. Por ejemplo, cuando se fabrican las puertas de un coche existe un valor objetivo para la anchura de la puerta. Para ser competitivo, las anchuras de las puertas producidas deben estar sistemáticamente próximas a ese valor. Los expertos citados dicen que el punto clave para producir unidades próximas al valor objetivo consiste en asegurar que la varianza de las unidades producidas sea mínima. Es decir, cuando se ajusta un proceso de producción de forma que las unidades producidas tengan valores con una varianza pequeña, se ha conseguido la parte más difícil para producir con regularidad unidades con unos valores que se encuentren próximos al valor objetivo.

La experiencia ha demostrado a esos expertos que, cuando se consigue que los valores de los elementos producidos tengan una varianza muy pequeña, es relativamente sencillo ajustar el proceso para que el valor medio de las unidades producidas esté próximo al valor objetivo. (Por analogía, dichos expertos dicen que si se quiere producir un rifle que permita al cazador alcanzar consistentemente un blanco, lo primero que se debe hacer es esforzarse en construir un rifle que sea extremadamente estable, es decir, que obtenga el mismo resultado cuando se apunta en la misma dirección; después, se deberá entrenar al usuario para que dispare en concordancia.)

Problemas

1. Se ha llevado a cabo un muestreo para averiguar la variación del número de horas semanales que trabajan los profesores universitarios. A partir de una muestra de 10 profesores se obtuvieron los siguientes datos:

   48, 22, 19, 65, 72, 37, 55, 60, 49, 28

   Utilícelos para estimar la desviación típica poblacional del número de horas que trabajan semanalmente los profesores de universidad.

2. Los siguientes datos muestran las anchuras (en pulgadas) de las muescas de nueve elementos de duraluminio que constituirán los bloques finales de las alas de un avión.

   8,751, 8,744, 8,749, 8,750, 8,752, 8,749, 8,764, 8,746, 8,753

   Estime la media y la desviación típica de las anchuras de las muescas del conjunto total de elementos de duraluminio producidos.

3. Los datos muestrales siguientes se refieren a la producción diaria (en toneladas) de una factoría química. Utilícelos para estimar la media y la varianza de la producción diaria.

   776, 810, 790, 788, 822, 806, 795, 807, 812, 791

4. La consistencia es muy importante en la producción de bolas de béisbol, ya que dichas bolas no pueden ser ni demasiado rápidas ni demasiado lentas. Para probar las bolas, se dejan caer desde un determinado nivel y, después, se mide la altura del bote. Si los siguientes estadísticos se han obtenido a partir de una muestra de 30 bolas

   $$\sum_{i=1}^{30} X_i = 52{,}1 \quad \sum_{i=1}^{30} X_i^2 = 136{,}2$$

   Estime la desviación típica de las alturas del bote para el total de las bolas de béisbol producidas. *Sugerencia*: Recuerde la identidad

   $$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

5. Utilice los datos del problema 1 de la sección 8.2 para estimar la desviación típica de los pesos de los corredores del maratón de Boston de 2004.

   Los problemas 6, 7 y 8 se refieren a los siguientes datos muestrales:

   104, 110, 114, 97, 105, 113, 106, 101, 100, 107

6. Estime la media poblacional μ y la varianza poblacional σ^2.

7. Supongamos que se sabe que la media de una población es 104. Estime la varianza poblacional.

8. Supongamos que se sabe que la media de una población es 106. Estime la desviación típica de la población.

9. Utilice los datos del problema 8 de la sección 8.2 para estimar la desviación típica de los tiempos de espera de los pacientes de la consulta médica del hospital.

10. Un fabricante de muebles decide probar una muestra de sillas resistentes al fuego para saber el calor que pueden soportar antes de que se inflamen. Se ha seleccionado una muestra de siete sillas y, una a una, se han colocado en una sala de pruebas. Una vez que se ha colocado una silla determinada se va aumentando grado a grado la temperatura de la sala de pruebas hasta que la silla se inflama. Supongamos que las temperaturas de inflamación (en grados Fahrenheit) de las siete sillas fueron las siguientes:

    458, 440, 482, 455, 491, 477, 446

    (a) Estime la temperatura media de inflamación para ese tipo de sillas.

    (b) Estime la desviación típica de las temperaturas de inflamación para dichas sillas.

11. Utilice los datos del problema 9 de la sección 8.2 para estimar la desviación típica de los tamaños de los hogares unifamiliares de la ciudad en cuestión.

12. Supongamos que la presión sanguínea sistólica de un minero se distribuye normalmente. Supongamos, además, que se han medido las presiones sanguíneas de los 13 mineros de una muestra aleatoria, y los resultados son:

    129, 134, 142, 114, 120, 116, 133, 142, 138, 148, 129, 133, 141

    (a) Estime la media de las presiones sanguíneas sistólicas de la población total de mineros.

    (b) Estime la desviación típica de las presiones citadas.

    (c) Utilice los estimadores de los apartados (a) y (b), junto con el hecho de que las presiones sanguíneas se distribuyen normalmente, para obtener un estimador de la proporción poblacional de mineros cuyas presiones sanguíneas sobrepasen 150.

13. El modelo lineal de paseo aleatorio para los precios diarios de una acción o de un determinado bien supone que las diferencias de los precios entre los distintos pares de días contiguos, en una sucesión de ellos, constituye una muestra aleatoria procedente de una población normal. Los siguientes 20 datos representan los precios de cierre en el mercado de crudo de Nueva York en 20 días laborables consecutivos del año 1994. Si se supone el modelo lineal de paseo aleatorio, utilice estos datos para estimar la media y la desviación típica de la distribución poblacional de las diferencias entre los precios de dos días consecutivos. (Observe que con los datos disponibles se pueden obtener 19 valores de la distribución citada; el primero es 17,60 − 17,50 = 0,10; el segundo, 17,81 − 17,60 = 0,21, y así sucesivamente.)

    17,50, 17,60, 17,81, 17,67, 17,53, 17,39, 17,12, 16,71, 16,70, 16,83,
    17,21, 17,24, 17,22, 17,67, 17,83, 17,67, 17,55, 17,68, 17,98, 18,39

14. Debido a la falta de precisión de la escala utilizada al pesar un pescado, el peso medido sigue una normal con media igual al peso verdadero y con desviación típica 0,1 gramos. Se ha seleccionado una muestra de 12 pescados *diferentes*, cuyos pesos resultaron ser los siguientes:

    5,5, 6,2, 5,8, 5,7, 6,0, 6,2, 5,9, 5,8, 6,1, 6,0, 5,7, 5,6

    Estime la desviación típica poblacional de los verdaderos pesos de los pescados.

    Sugerencia: Observe en primer lugar que, debido a los errores con el sistema para pesar, cada peso medido no coincide con peso real del pescado, sino que, por el contrario, es la suma del peso real más el error cometido. Estos errores son variables aleatorias, independientes de los pesos verdaderos, con una media de 0 y con una desviación típica de 0,1. Así pues,

    Dato = peso verdadero + error

y, por tanto,

$$\text{Var (dato)} = \text{Var (peso verdadero)} + \text{Var (error)}$$

Para determinar la varianza de los pesos verdaderos estime la varianza de los datos.

8.5 Estimadores por intervalo para la media de una población normal con una varianza conocida

Cuando se estima un parámetro mediante un estimador puntual no se puede esperar que el estimador resultante sea exactamente igual al parámetro, sino que esté "próximo" a él. Sin embargo, en ocasiones, uno quiere ser más concreto y busca un intervalo construido alrededor del estimador puntual, para el cual tengamos una elevada confianza en que el parámetro esté contenido en dicho intervalo. Éste último recibe el nombre de *estimador por intervalo*.

Definición

Un *estimador por intervalo* de un parámetro poblacional es un intervalo para el que se predice que el parámetro está contenido en él. La *confianza* que se da al intervalo es la probabilidad de que el intervalo contenga al parámetro.

Para obtener un estimador por intervalo para un parámetro de la población se utiliza la distribución de probabilidad del estimador puntual del parámetro. Veamos cómo se puede obtener un estimador por intervalo para la media de una población normal cuando se conoce la desviación típica poblacional.

Sea $X_1, \ldots, X_n$ una muestra de tamaño n procedente de una población normal con desviación típica conocida σ, y supongamos que se va a utilizar esta muestra para obtener un estimador por intervalo, con un 95% de confianza, para la media μ de la población. Para determinar ese intervalo se partirá de la media muestral $\overline{X}$, que es un estimador puntual de μ. Se tendrá en cuenta que $\overline{X}$ sigue una normal con una media μ y con una desviación típica $\sigma/\sqrt{n}$, lo que implica que la variable estandarizada

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \sqrt{n}\,\frac{\overline{X} - \mu}{\sigma}$$

sigue una distribución normal estándar. Ahora bien, dado que $z_{0,025} = 1{,}96$, se tiene que, en un 95% de los casos, el valor absoluto de Z es menor o igual que 1,96 (véase la figura 8.1).

Así pues, se puede escribir

$$P\left\{\frac{\sqrt{n}}{\sigma}|\overline{X} - \mu| \leq 1{,}96\right\} = 0{,}95$$

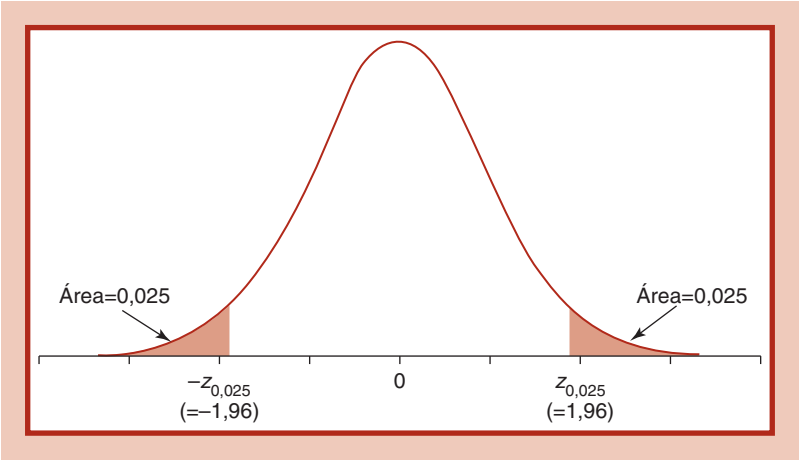


Figura 8.1 $P\{|Z| \leq 1{,}96\} = P\{-1{,}96 \leq Z \leq 1{,}96\} = 0{,}95$.

Si se multiplican los dos miembros de la desigualdad por $\sigma/\sqrt{n}$, se ve que la ecuación anterior es equivalente a

$$P\left\{|\overline{X} - \mu| \leq 1{,}96\,\frac{\sigma}{\sqrt{n}}\right\} = 0{,}95$$

De ello se desprende que, con una probabilidad del 95%, μ y $\overline{X}$ distan entre sí como máximo $1{,}96\sigma/\sqrt{n}$. Esto equivale a que

$$P\left\{\overline{X} - 1{,}96\,\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 1{,}96\,\frac{\sigma}{\sqrt{n}}\right\} = 0{,}95$$

Es decir, con un 95% de confianza, la media poblacional se encuentra en el intervalo $\overline{X} \pm 1{,}96\sigma/\sqrt{n}$.

El intervalo con extremos $\overline{X} - 1{,}96\sigma/\sqrt{n}$ y $\overline{X} + 1{,}96\sigma/\sqrt{n}$ se dice que es un *estimador por intervalo, con un 95% de confianza,* para la media poblacional μ. Si el valor observado de $\overline{X}$ es $\overline{x}$, el intervalo con extremos $\overline{x} \pm 1{,}96\sigma/\sqrt{n}$ es el *valor estimado del estimador por intervalo de μ, con un 95% de confianza.*

Ejemplo 8.5 Supongamos que una señal de intensidad μ emitida desde el punto A se registra en un punto B con una intensidad que se distribuye según una normal de media μ y desviación típica 3. Esto es, debido al "ruido", la intensidad registrada difiere de intensidad real en una cantidad que se distribuye normalmente con media 0 y desviación típica 3.

Para reducir este error, la misma señal se registra independientemente 10 veces. Si los sucesivos valores registrados son

$$17, 21, 20, 18, 19, 22, 20, 21, 16, 19$$

construya un intervalo al 95% de confianza para la intensidad real μ.

Solución El valor de la media muestral es

$$\frac{17 + 21 + 20 + 18 + 19 + 22 + 20 + 21 + 16 + 19}{10} = 19{,}3$$

Puesto que $\sigma = 3$ se tiene que el estimador por intervalo al 95% de confianza para μ viene dado por

$$19{,}3 \pm 1{,}96 \frac{3}{\sqrt{10}} = 19{,}3 \pm 1{,}86$$

Es decir, se puede asegurar, con un 95% de confianza, que la intensidad real de la señal está comprendida entre 17,44 y 21,16. En la figura 8.2 se muestra un gráfico de este intervalo. ■

También se pueden obtener estimadores por intervalo con niveles de confianza distintos de 0,95. Observe que para cualquier valor α comprendido entre 0 y 1, la probabilidad de que una normal estándar esté comprendida entre $-z_{\alpha/2}$ y $z_{\alpha/2}$ es igual a $1 - \alpha$ (véase la figura 8.3). De aquí se desprende que

$$P\left\{\frac{\sqrt{n}}{\sigma}|\overline{X} - \mu| \leq z_{\alpha/2}\right\} = 1 - \alpha$$

Con un razonamiento similar al empleado anteriormente cuando $\alpha = 0{,}05$ ($z_{0{,}025} = 1{,}96$), se puede demostrar que, con probabilidad $1 - \alpha$, μ está en el intervalo $\overline{X} \pm z_{\alpha/2}\sigma/\sqrt{n}$.

El intervalo $\overline{X} \pm z_{\alpha/2}\sigma/\sqrt{n}$. se denomina *estimador por intervalo, al 100(1 − α)% de confianza,* para la media de la población.

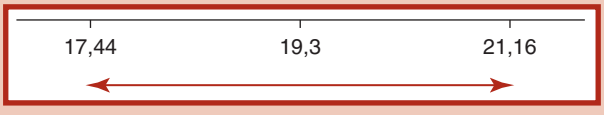


Figura 8.2 Estimador por intervalo de confianza de μ en el problema 8.5.

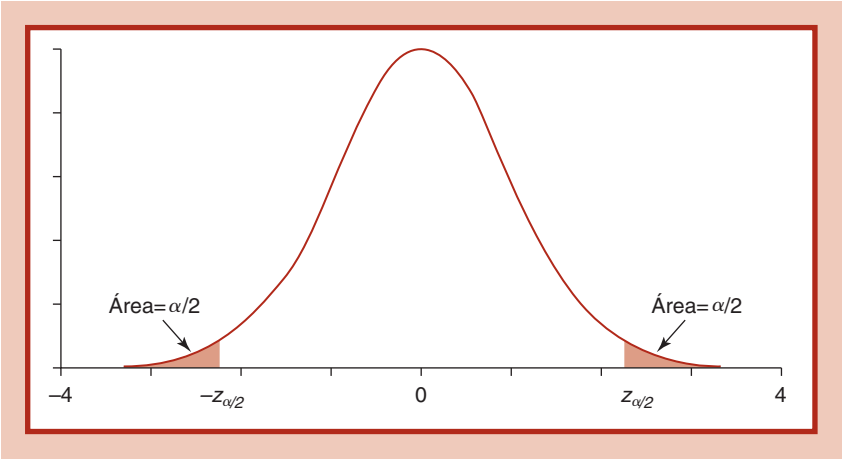


Figura 8.3 $P\{|Z| \leq z_{\alpha/2}\} = P\{-z_{\alpha/2} \leq Z \leq z_{\alpha/2}\} = 1 - \alpha$.

Tabla 8.1 Percentiles correspondientes a distintos niveles de confianza

Nivel de confianza $100(1-\alpha)$	Valor de α correspondiente	Valor de $z_{\alpha/2}$
90	0,10	$z_{0,05} = 1,645$
95	0,05	$z_{0,025} = 1,960$
99	0,01	$z_{0,005} = 2,576$

La tabla 8.1 muestra los valores de $z_{\alpha/2}$ necesarios para obtener los estimadores por intervalo para μ, con niveles de confianza del 90, 95 y 99 por ciento, respectivamente.

Ejemplo 8.6 Calcule, para los datos del ejemplo 8.5:

(a) Un estimador por intervalo para μ, con una confianza del 90%.

(b) Un estimador por intervalo para μ, con una confianza del 99%.

Solución Se debe construir un estimador por intervalo para μ, a confianza de $100(1-\alpha)\%$, siendo $\alpha = 0,10$ en el apartado (a) y $\alpha = 0,01$ en el apartado (b). Ahora bien,

$$z_{0,05} = 1,645 \qquad \text{y} \qquad z_{0,005} = 2,576$$

por consiguiente, el estimador por intervalo pedido, a confianza del 90%, es

$$\overline{X} \pm 1,645 \frac{\sigma}{\sqrt{n}}$$

mientras que el estimador por intervalo pedido, a confianza del 99%, es

$$\overline{X} \pm 2,576 \frac{\sigma}{\sqrt{n}}$$

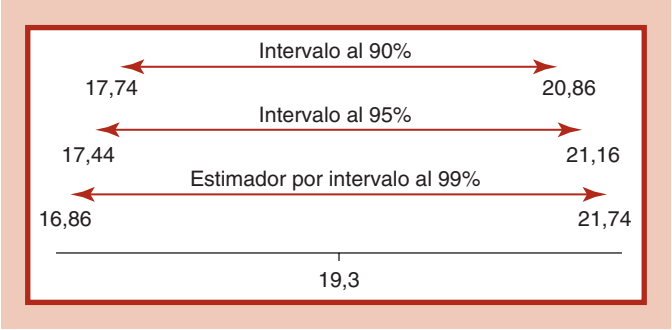


Figura 8.4 Estimadores por intervalo, con confianzas del 90, el 95 y el 99 por ciento.

Para los datos del ejemplo 8.5, $n = 10$, $\overline{X} = 19{,}3$, y $\sigma = 3$. Por consiguiente, los valores estimados del estimador por intervalo, a confianzas del 90% y el 99%, para μ son, respectivamente,

$$19{,}3 \pm 1{,}645 \frac{3}{\sqrt{10}} = 19{,}3 \pm 1{,}56$$

y

$$19{,}3 \pm 2{,}576 \frac{3}{\sqrt{10}} = 19{,}3 \pm 2{,}44$$

La figura 8.4 muestra los estimadores por intervalo para μ, a confianzas del 90, 95 y 99 por ciento. Observe que cuanto mayor es el coeficiente de confianza $100(1 - \alpha)$, mayor es la longitud del intervalo. Esto tiene sentido ya que, si se desea aumentar la certidumbre de que el parámetro esté contenido en un determinado intervalo, claramente éste se tendrá que alargar. ■

En ocasiones, se está interesado en obtener un intervalo con una confianza del $100(1 - \alpha)$ por ciento cuya longitud sea menor o igual que un valor prefijado, y el problema consiste en elegir el tamaño muestral adecuado que lo permita. Por ejemplo, supongamos que se desea obtener un intervalo, con una longitud como máximo b, que contenga la media poblacional con un 95% de certeza. ¿Cuál debe ser el tamaño muestral? Para responder a esta pregunta observe que, dado que $z_{0{,}025} = 1{,}96$, el intervalo al 95% de confianza para μ basado en una muestra de tamaño n, es (véase la figura 8.5)

$$\overline{X} \pm 1{,}96 \frac{\sigma}{\sqrt{n}}$$

Puesto que la longitud de este intervalo es

$$\text{Longitud del intervalo} = 2(1{,}96) \frac{\sigma}{\sqrt{n}} = 3{,}92 \frac{\sigma}{\sqrt{n}}$$

se debe elegir n de forma que

$$\frac{3{,}92\sigma}{\sqrt{n}} \leq b$$

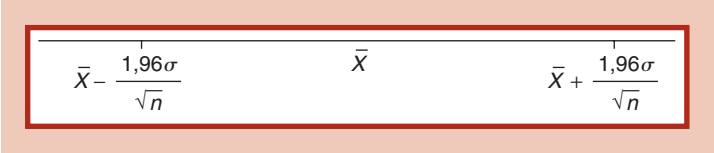


Figura 8.5 Intervalo para μ, a confianza del 95%.

o, equivalentemente,

$$\sqrt{n} \geq \frac{3{,}92\sigma}{b}$$

Si se elevan al cuadrado los dos miembros de esta desigualdad se obtiene que el tamaño muestral se debe elegir de modo que

$$n \geq \left(\frac{3{,}92\sigma}{b}\right)^2$$

Ejemplo 8.7 Supongamos que la desviación típica de una población es $\sigma = 2$. Si se desea obtener un estimador por intervalo para μ, a confianza del 95%, cuya longitud sea menor o igual que $b = 0{,}01$, ¿qué tamaño muestral se necesita?

Solución Se debe elegir un tamaño muestral n tal que

$$n \geq \left(\frac{3{,}92 \times 2}{0{,}1}\right)^2 = (78{,}4)^2 = 6146{,}6$$

Esto es, como mínimo se necesita un tamaño muestral de 6147 unidades. ■

El procedimiento para calcular el tamaño muestral requerido para obtener un intervalo para μ, con una confianza del $100(1 - \alpha)\%$ y con una longitud menor o igual que b, es similar al explicado cuando $\alpha = 0{,}05$. El resultado es el siguiente:

Determinación del tamaño muestral necesario

La longitud del estimador por intervalo para la media poblacional, a confianza del $100(1 - \alpha)\%$, resulta ser menor o igual que b si el tamaño muestral n verifica que

$$n \geq \left(\frac{2z_{\alpha/2}\sigma}{b}\right)^2$$

Una vez elegido n, el estimador por intervalo de confianza de longitud menor que b será

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

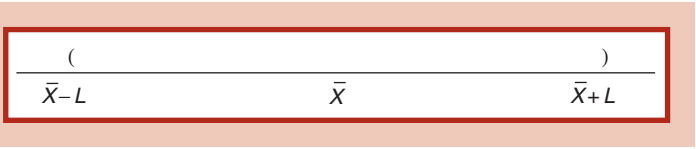


Figura 8.6 Intervalo de confianza centrado en $\bar{X}$. Si la longitud del intervalo es $2L$, la distancia de $\bar{X}$ a cualquier punto del intervalo es menor o igual que L.

Ejemplo 8.8 Se sabe por la experiencia que los pesos de los salmones de una piscifactoría se distribuyen normalmente con una media que varía de una temporada a otra y con una desviación típica que se mantiene igual a 0,3 libras. Si, con una confianza del 90%, se pretende obtener un estimador por intervalo para el peso medio de los salmones de una temporada, de modo que $\bar{X}$ difiera de μ menos de $\pm 0{,}1$ libras, ¿qué tamaño muestral se necesita? ¿Y si se pretendiera que el intervalo tuviera una confianza del 99%?

Solución Puesto que el estimador por intervalo, a confianza del 90%, obtenido con una muestra de tamaño n es $\bar{X} \pm 1{,}645\sigma/\sqrt{n}$, se tiene que, con una confianza del 90%, el estimador puntual diferirá de μ menos de $\pm 0{,}1$ libras, siempre que la longitud de ese intervalo de confianza sea menor o igual que 0,2 (véase la figura 8.6). Por consiguiente, de lo anterior se desprende que se debe elegir n de forma que

$$n \geq \left(\frac{2 \times 1{,}645 \times 0{,}3}{0{,}2}\right)^2 = 24{,}35$$

Esto es, como mínimo se requiere un tamaño muestral de 25 unidades.

Si, por otra parte, se quisiera tener una certeza del 99% de que $\bar{X}$ difiera de la media real en menos de $\pm 0{,}1$ libras, se debe tener en cuenta que $z_{0{,}005} = 2{,}576$ y, por tanto, el tamaño muestral debe satisfacer

$$n \geq \left(\frac{2 \times 2{,}576 \times 0{,}3}{0{,}2}\right)^2 = 59{,}72$$

Esto es, el mínimo tamaño muestral que se precisa es 60. ■

Para obtener el estimador por intervalo de la media de una normal cuya varianza se conoce se ha utilizado el hecho de que $\bar{X}$ se distribuye según una normal con media μ y desviación típica $\sigma/\sqrt{n}$. Sin embargo, por el teorema central del límite, lo anterior continúa siendo aproximadamente cierto para la media de cualquier distribución poblacional si el tamaño muestral n es suficientemente grande ($n \geq 30$ en la mayoría de los casos es suficiente). Como resultado se puede utilizar el intervalo $\bar{X} \pm z_{\alpha/2}\sigma/\sqrt{n}$ como estimador por intervalo, a confianza de $100(1 - \alpha)\%$, para la media de cualquier población, siempre que el tamaño muestral sea lo suficientemente grande como para que se pueda aplicar el teorema central del límite.

Ejemplo 8.9 Para estimar μ, el contenido medio de nicotina en los cigarrillos de una nueva marca que ha salido al mercado, se han seleccionado aleatoriamente 44 cigarrillos de dicha marca y se han medido sus contenidos en nicotina.

(a) Si el contenido medio en nicotina de la muestra fue de 1,74 miligramos, ¿cuál será el estimador por intervalo de μ, al 95% de confianza?

(b) ¿Qué tamaño muestral se necesita para que la longitud del intervalo, al 95% de confianza, sea menor o igual que 0,3 miligramos?

Suponga que se conoce por la experiencia que la desviación típica de los contenidos de nicotina por cigarrillo es igual a 0,7 miligramos.

Solución

(a) Dado que 44 es un tamaño muestral suficientemente grande, no es necesario que la distribución poblacional sea normal para poder asegurar que

$$\overline{X} \pm z_{0,025} \frac{\sigma}{\sqrt{n}}$$

es un estimador por intervalo, al 95% de confianza, para la media poblacional. En nuestro caso, el intervalo estimado es

$$1,74 \pm \frac{1,96(0,7)}{\sqrt{44}} = 1,74 \pm 0,207$$

Esto es, se puede asegurar con un 95% de confianza que el contenido medio de nicotina por cigarrillo está comprendido entre 1,533 y 1,947 miligramos.

(b) La longitud del estimador por intervalo al 95% de confianza será menor o igual que 0,3 si el tamaño muestral n verifica que

$$n \geq \left(\frac{2 \times 1,96 \times 0,7}{0,3} \right)^2 = 83,7$$

Esto es, se necesita una muestra de tamaño 84 como mínimo. ■

8.5.1 Cotas superior e inferior de confianza

En ocasiones se está interesado en asegurar con una confianza dada que la media de la población es mayor que un determinado valor. Para obtener la citada *cota inferior de confianza* para la media poblacional, de nuevo se tendrá en cuenta que

$$Z = \sqrt{n} \frac{\overline{X} - \mu}{\sigma}$$

sigue una distribución normal estándar. En consecuencia, se verifica que (véase la figura 8.7)

$$P\left\{ \sqrt{n} \frac{\overline{X} - \mu}{\sigma} < z_\alpha \right\} = 1 - \alpha$$

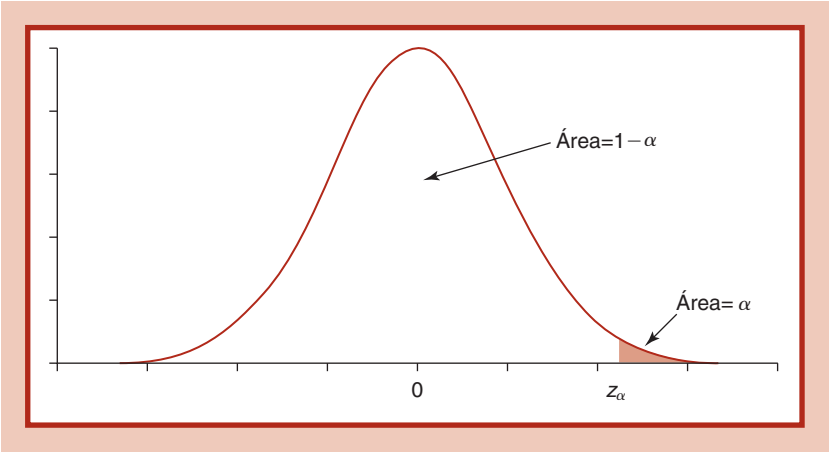


Figura 8.7 $P\{Z \leq z_a\} = 1 - \alpha$.

que puede reescribirse como

$$P\left\{\mu > \overline{X} - z_\alpha \frac{\sigma}{\sqrt{n}}\right\} = 1 - \alpha$$

De esta ecuación se concluye lo siguiente:

Una cota inferior, al $100(1 - \alpha)\%$ de confianza, para la media poblacional μ viene dada por

$$\overline{X} - z_\alpha \frac{\sigma}{\sqrt{n}}$$

Esto es, con un $100(1 - \alpha)\%$ de confianza, se puede asegurar que

$$\mu > \overline{X} - z_\alpha \frac{\sigma}{\sqrt{n}}$$

Ejemplo 8.10 Supongamos que en el ejemplo 8.8 se desea obtener un valor que sea menor que el peso medio de los salmones, con un 95% de confianza. Si el peso medio de la muestra de 50 salmones resultó ser de 5,6 libras, calcule ese valor.

Solución Se pide que calculemos una cota inferior para μ, al 95% de confianza. Del desarrollo anterior se desprende que dicha cota viene dada por

$$\overline{X} - z_{0,05} \frac{\sigma}{\sqrt{n}}$$

Puesto que $z_{0,05} = 1,645$, $\sigma = 0,3$, $n = 50$ y $\overline{X} = 5,6$, la cota inferior de confianza coincide con

$$5,6 - 1,645\frac{0,3}{\sqrt{50}} = 5,530$$

Esto es, se puede asegurar, con una confianza del 95%, que el peso medio de los salmones es mayor que 5,530 libras. ■

De forma similar se puede obtener una cota superior para μ, al $100(1 - \alpha)\%$ de confianza. Ésta se indica a continuación.

Una cota superior para la media poblacional μ, al $100(1 - \alpha)\%$ de confianza, viene dada por

$$\overline{X} + z_a\frac{\sigma}{\sqrt{n}}$$

Esto es, con un $100(1 - \alpha)$ por ciento de confianza, se puede asegurar que

$$\mu < \overline{X} + z_\alpha\frac{\sigma}{\sqrt{n}}$$

Ejemplo 8.11 En el ejemplo 8.9, calcule una cota superior para μ, al 95% de confianza.

Solución Una cota superior para μ, al 95% de confianza, viene dada por

$$\overline{X} + z_{0,05}\frac{\sigma}{\sqrt{n}} = 1,74 + 1,645\frac{0,7}{\sqrt{44}} = 1,914$$

Esto es, se puede asegurar, con un 95% de confianza, que el contenido medio de nicotina es inferior a 1,914 miligramos. ■

Problemas

1. Un peso eléctrico realiza mediciones iguales al peso verdadero más un error aleatorio que se distribuye como una normal con media 0 y desviación típica $\sigma = 0,1$ onzas. Supongamos que los resultados de 5 pesadas sucesivas de un mismo objeto fueron los siguientes: 3,142, 3,163, 3,155, 3,150 y 3,141.

   (a) Obtenga un estimador por intervalo para el peso verdadero, al 95% de confianza.

   (b) Calcule un estimador por intervalo para el peso verdadero, al 99% de confianza.

2. Un administrador de un hospital mantiene que, tras realizar un estudio estadístico, se puede asegurar lo siguiente: “Con una certeza del 90%, el peso medio de los niños

recién nacidos en el hospital se encuentra entre 6,6 y 7,2 libras." ¿Cómo se debe interpretar esta sentencia?

3. Se midió la concentración de bifenil policlorinado (PCB) en un pez que había sido pescado en el lago Michigan mediante una técnica de la que se sabe que produce un error que se distribuye según una normal con desviación típica 0,08 partes por millón. Si los resultados de 10 medidas independientes realizadas sobre el pez fueron

   11,2, 12,4, 10,8, 11,6, 12,5, 10,1, 11,0, 12,2, 12,4, 10,6

   obtenga un estimador por intervalo, al 95% de confianza, para el nivel de PCB de ese pez.

4. Supongamos que, en el problema 3, se han realizado 40 medidas, a partir de las cuales se ha obtenido el mismo valor medio especificado en dicho problema. Calcule el estimador por intervalo, al 95% de confianza, para el nivel de PCB del pez analizado.

5. Se sabe que la duración del tubo de una determinada marca de televisión se distribuye normalmente con una desviación típica de 400 horas. Supongamos que la duración media de una muestra aleatoria de 20 tubos fue de 9000 horas. Obtenga un estimador por intervalo para la duración media de los tubos de televisión citados, con un nivel de confianza del:

   (a) 95%

   (b) 99%

6. Una compañía de ingeniería fabrica un componente de un cohete espacial cuya longitud de vida se distribuye normalmente con una desviación típica de 3,4 horas. Si la vida media de una muestra de 9 de los citados componentes fue de 10,8 horas, calcule un estimador por intervalo para la vida media de dichos componentes, con un nivel de confianza del:

   (a) 95%

   (b) 99%

7. La desviación típica de las puntuaciones obtenidas en un test es de 11,3 puntos. La puntuación media de las calificaciones obtenidas por los estudiantes de una muestra de tamaño 81 fue de 74,6 puntos. Obtenga un estimador por intervalo para la calificación media del total de estudiantes, al nivel de confianza del 90%.

8. En el problema 7 supongamos que la calificación media muestral fue de 74,6 puntos, con una muestra de tamaño 324. Obtenga, de nuevo, un estimador por intervalo, al 90% de confianza.

9. Se sabe que la desviación típica de las duraciones de cierto tipo de bombillas es igual a 100 horas. La duración media de una muestra de 169 de las bombillas citadas fue de 1350 horas. Calcule un estimador por intervalo para la duración media del total de bombillas de ese tipo, con un nivel de confianza del:

   (a) 90%

   (b) 95%

   (c) 99%

10. La vida media de una muestra aleatoria de 10 cubiertas de cierta marca fue de 28 400 millas. Si se sabe que las duraciones de esas cubiertas se distribuyen normalmente con una desviación típica de 3300 millas, obtenga un estimador por intervalo para la duración media de las cubiertas de dicha marca, al 95% de confianza.

11. En el problema 10, con un nivel de confianza del 99%, ¿cuál debe ser el tamaño de la muestra para obtener un estimador por intervalo con una longitud inferior al allí calculado?

12. Un estudio piloto ha revelado que la desviación típica de los salarios mensuales de los trabajadores de la industria química es de 180 dólares. Con un nivel de confianza del 90%, ¿cuál debe ser el tamaño muestral para que el estimador puntual del salario medio poblacional difiera del salario medio real en ± 20 dólares?

13. Repita el problema 12 si la confianza que se pretende conseguir es del 95%.

14. Una gestora de la política de admisión de una universidad desea conocer la calificación media que los estudiantes que acaban de entrar en la universidad han conseguido en un test de aptitud previo. En lugar de consultar los archivos al completo decide seleccionar una muestra aleatoria de estudiantes. Si se sabe que las calificaciones de los estudiantes se distribuyen normalmente con una desviación típica de 70 puntos, ¿cuál debe ser el tamaño de la muestra si la gestora pretende obtener un estimador por intervalo que tenga una longitud como máximo de 4 puntos?

15. En el problema 7, para la calificación media del test, calcule:

    (a) Una cota inferior, al 90% de confianza.

    (b) Una cota inferior, al 95% de confianza.

    (c) Una cota superior, al 95% de confianza.

    (d) Una cota superior, al 99% de confianza.

16. Los siguientes datos muestrales proceden de una población normal con desviación típica 3:

    5, 4, 8, 12, 11, 7, 14, 12, 15, 10

    (a) Calcule un valor que, a 95% de confianza, supere la media poblacional.

    (b) Obtenga un valor que, con una confianza del 99%, sea inferior a la media poblacional.

17. Suponga, con los datos del problema 10, que el fabricante de cubiertas anuncia lo siguiente: “Con una certeza del 95%, la vida media de las cubiertas está por encima de 26 000 millas”. ¿Es falso este anuncio?

8.6 Estimadores por intervalo para la media de una población normal con varianza desconocida

Supongamos que se ha extraído una muestra $X_1, \ldots, X_n$ procedente de una población normal con media μ y desviación típica σ, ambas desconocidas, con cuyos datos se pretende obtener un estimador por intervalo para la media poblacional μ.

Para empezar, recordemos cómo se obtenía un estimador por intervalo para μ cuando se suponía que σ era conocida. Se partía del hecho de que Z, la versión estandarizada del estimador puntual $\overline{X}$, que viene dada por

$$Z = \sqrt{n}\frac{\overline{X} - \mu}{\sigma}$$

sigue una distribución normal estándar. Puesto que se supone ahora que σ es desconocida, parece natural sustituirla por su estimador S, la desviación típica muestral, y, por consiguiente, basar nuestro intervalo de confianza sobre la variable T_{n-1} dada por

$$T_{n-1} = \sqrt{n}\frac{\overline{X} - \mu}{S}$$

La variable aleatoria T_{n-1} definida arriba se dice que es una variable aleatoria t con $n - 1$ grados de libertad.

La variable aleatoria

$$T_{n-1} = \sqrt{n}\frac{\overline{X} - \mu}{S}$$

se dice que es una *variable aleatoria t con n − 1 grados de libertad.*

La razón por la que T_{n-1} tiene $n - 1$ grados de libertad es que la varianza muestral S^2, utilizada para estimar σ^2, se distribuye, tras multiplicarse por $(n - 1)/\sigma^2$, como una chi-cuadrado con $n - 1$ grados de libertad (véase la sección 7.6).

La función de densidad de una variable aleatoria t, al igual que la de una normal estándar, es simétrica respecto de cero. Su apariencia es muy similar a la densidad de la normal estándar, aunque su dispersión es mayor ya que "sus colas son más pesadas". A medida de que los grados de libertad aumentan, la densidad de la t se aproxima más a la de la normal estándar. La figura 8.8 muestra gráficamente las funciones de densidad de las variables aleatorias t para una gran variedad de grados de libertad.

El valor $t_{n\,\alpha}$ se define de forma que

$$P\{T_n > t_{n,\alpha}\} = \alpha$$

siendo T_n una variable aleatoria t con n grados de libertad (véase la figura 8.9).

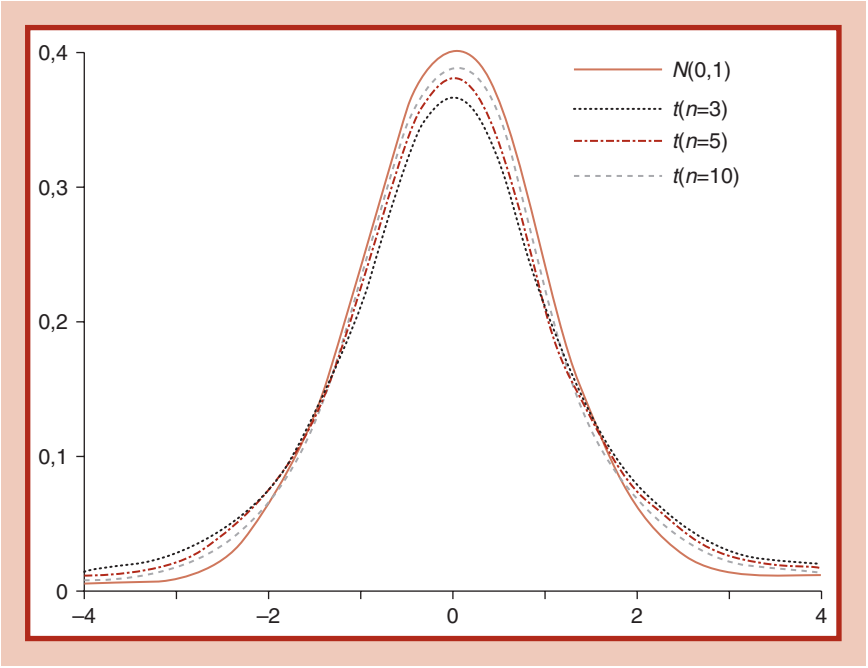


Figura 8.8 Distribuciones normal estándar y t.

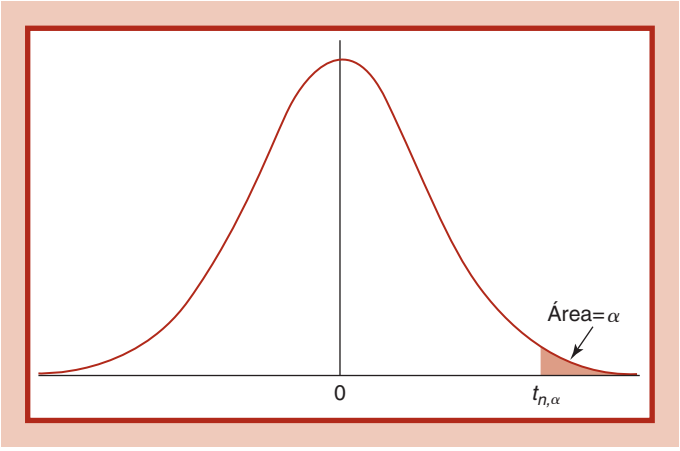


Figura 8.9 Percentil de la densidad: $P\{T_n > t_{n,\alpha}\} = \alpha$.

Puesto que $P\{T_n < t_{n,\alpha}\} = 1 - \alpha$, se tiene que $t_{n,\alpha}$ es el percentil de orden $100(1 - \alpha)\%$ de la distribución t con n grados de libertad. Por ejemplo, $P\{T_n < t_{n,\,0,05}\} = 0{,}95$, lo que indica que, en un 95% de los casos, una variable aleatoria t con n grados de libertad toma valores menores que $t_{n,0,05}$. El valor $t_{n,\alpha}$ es análogo a z_α de la distribución normal estándar.

En la tabla D.2 del Apéndice D se muestran los valores de $t_{n,\alpha}$ para distintas cuantías de n y α. Adicionalmente, el Programa 8-1 computa los valores de a percentiles. También se puede utilizar el Programa 8-2 para obtener las probabilidades asociadas a una variable aleatoria t.

Ejemplo 8.12 Indique el valor de $t_{8,\,0,05}$.

Solución Se puede obtener el valor de $t_{8,\,0,05}$ a partir de la tabla D.2. A continuación se muestra una parte de esta tabla.

Valores de $t_{n,\alpha}$

n	$\alpha = 0{,}10$	$\alpha = 0{,}05$ ↓	$\alpha = 0{,}025$
6	1,440	1,943	2,447
7	1,415	1,895	2,365
→8	1,397	1,860	2,306
9	1,383	1,833	2,262

Moviéndonos de arriba abajo en la columna $\alpha = 0{,}05$ hasta llegar a la fila $n = 8$, se encuentra que $t_{8,\,0,05} = 1{,}860$. ■

Por la simetría de la distribución t respecto de *cero,* se tiene (véase la figura 8.10) que

$$P\{|T_n| \le t_{n,\alpha/2}\} = P\{-t_{n,\alpha/2} \le T_n \le t_{n,\alpha/2}\} = 1 - \alpha$$

Puesto que $\sqrt{n}(\bar{X} - \mu)/S$ sigue una distribución t con $n - 1$ grados de libertad, se desprende de lo anterior que

$$P\left\{\sqrt{n}\,\frac{|\bar{X} - \mu|}{S} \le t_{n-1,\alpha/2}\right\} = 1 - \alpha$$

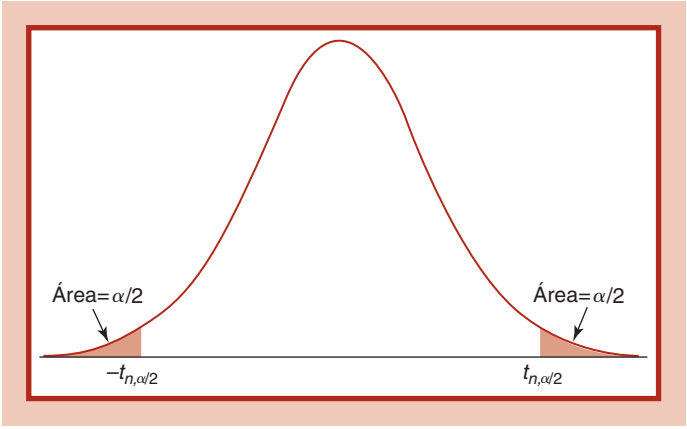


Figura 8.10 $P\{|T_n| \le t_{n,\alpha/2}\} = P\{-t_{n,\alpha/2} \le T_n \le t_{n,\alpha/2}\} = 1 - \alpha$.

Razonando exactamente igual que cuando σ era conocida se puede demostrar que la ecuación anterior es equivalente a

$$P\left\{\overline{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right\} = 1 - \alpha$$

Por consiguiente, se ha demostrado lo siguiente:

Un estimador por intervalo, al $100(1 - \alpha)\%$ de confianza, para la media poblacional μ viene dado por el intervalo

$$\overline{X} \pm t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}$$

El Programa 8-3 permite obtener el anterior estimador por intervalo, con una confianza dada, a partir de un conjunto de datos muestrales.

Ejemplo 8.13 La Agencia de Protección Medioambiental de Estados Unidos está analizando las cantidades de PCB, un producto tóxico, en la leche materna. Las cantidades de PCB (en partes por millón) encontrada en la leche de una muestra de 20 mujeres con hijos lactantes fueron las siguientes:

16, 0, 0, 2, 3, 6, 8, 2, 5, 0, 12, 10, 5, 7, 2, 3, 8, 17, 9, 1

Utilice estos datos para obtener un:

(a) intervalo de confianza al 95%,

(a) intervalo de confianza al 99%,

para la cantidad media de PCB en la leche de las madres con hijos lactantes.

Solución Tras cálculos sencillos se obtiene que la media muestral y la desviación típica muestral son

$$\overline{X} = 5{,}8 \qquad S = 5{,}085$$

Puesto que $100(1 - \alpha)$ es igual a 0,95 si $\alpha = 0{,}05$ y es igual a 0,99 si $\alpha = 0{,}01$, se deben encontrar los valores de $t_{19,\,0{,}025}$ y $t_{19,\,0{,}005}$. En la tabla D.2 se ve que

$$t_{19,\,0{,}025} = 2{,}093 \qquad t_{19,\,0{,}005} = 2{,}861$$

De aquí se desprende que el estimador por intervalo, al 95% de confianza, para μ es

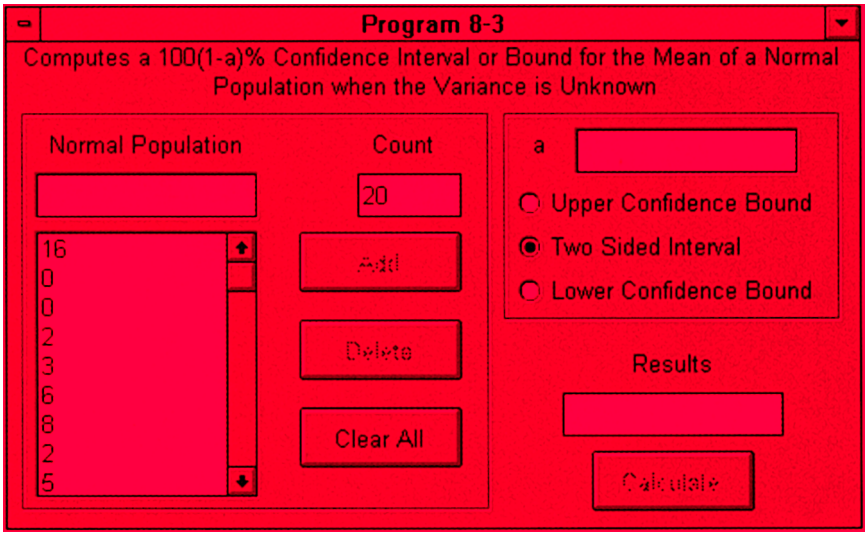
$$5{,}8 \pm 2{,}093\frac{5{,}085}{\sqrt{20}} = 5{,}8 \pm 2{,}38$$

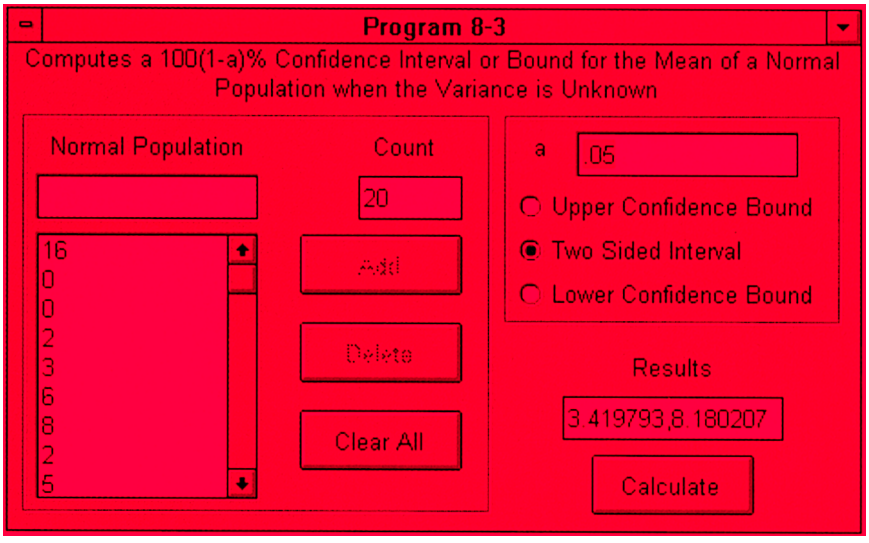
mientras que, al 99% de confianza, el intervalo es

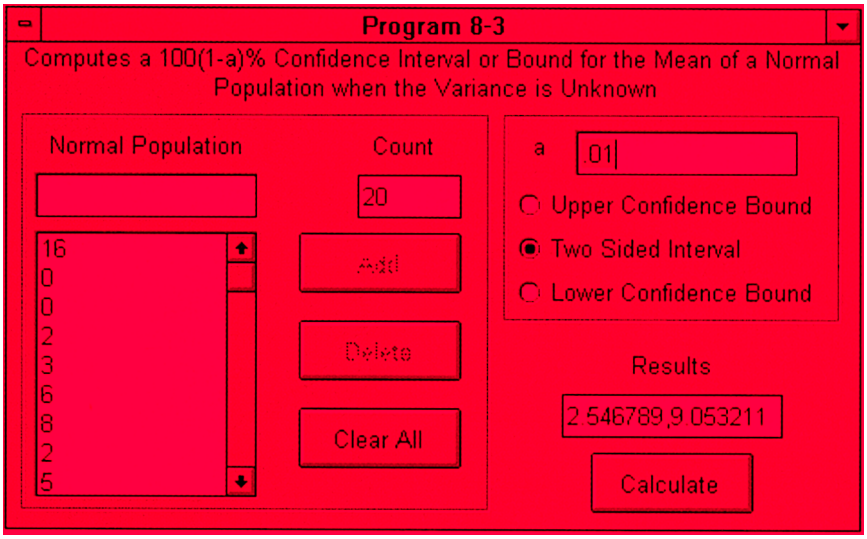
$$5{,}8 \pm 2{,}861 \frac{5{,}085}{\sqrt{20}} = 5{,}8 \pm 3{,}25$$

Esto es, con una confianza del 95%, la cantidad media de PCB en la leche materna está comprendida entre 3,42 y 8,18 partes por millón; mientras que, al 99% de confianza, la cantidad media citada está entre 2,55 y 9,05 partes por millón.

Este ejemplo se podría haber resuelto utilizando el Programa 8-3, con el que se obtiene lo siguiente.







8.6.1 Cotas inferior y superior de confianza

También se pueden calcular muy sencillamente las cotas inferior y superior de confianza para μ, obteniéndose los resultados siguientes:

La cota inferior para μ, al $100(1-\alpha)\%$ de confianza, viene dada por

$$\overline{X} - t_{n-1,\alpha}\frac{S}{\sqrt{n}}$$

Esto significa que, con una confianza del $100(1-\alpha)\%$, la media poblacional es mayor que

$$\overline{X} - t_{n-1,\alpha}\frac{S}{\sqrt{n}}$$

La cota superior para μ, al $100(1-\alpha)\%$ de confianza, viene dada por

$$\overline{X} + t_{n-1,\alpha}\frac{S}{\sqrt{n}}$$

Esto es, con una confianza del $100(1-\alpha)\%$, la media poblacional es menor que

$$\overline{X} + t_{n-1,\alpha}\frac{S}{\sqrt{n}}$$

Ejemplo 8.14 En el ejemplo 8.13, calcule:

(a) una cota superior, al 95% de confianza,

(b) una cota superior, al 99% de confianza,

para la cantidad media de PCB en la leche materna.

Solución En el ejemplo 8.13, el tamaño muestral era igual a 20, y los valores de la media muestral y de la desviación típica muestral resultaron ser

$$\overline{X} = 5{,}8 \quad S = 5{,}085$$

(a) De la tabla D.2 se obtiene que

$$t_{19,\,0{,}05} = 1{,}729$$

Así pues, la cota superior, al 95% de confianza, es

$$5{,}8 + 1{,}729\frac{5{,}085}{\sqrt{20}} = 7{,}77$$

Esto es, con una confianza del 95%, se puede asegurar que el nivel medio de PCB en la leche de las madres con hijos lactantes es menor que 7,77 partes por millón.

(b) De la tabla D.2,

$$t_{19,\,0{,}01} = 2{,}539$$

Por tanto, la cota inferior, al 99% de confianza, es

$$5{,}8 - 2{,}539\frac{5{,}085}{\sqrt{20}} = 2{,}91$$

así pues, con una confianza del 99%, se puede mantener que el nivel medio de PCB en la leche de las madres con hijos lactantes es mayor que 2,91 partes por millón. ■

El Programa 8-3 también permite computar las cotas superior e inferior con una confianza dada.

Problemas

1. El Centro Nacional de Estadísticas de Educación de Estados Unidos ha seleccionado una muestra aleatoria de 2000 estudiantes universitarios recién graduados, a quienes se les preguntó cuánto tiempo habían empleado en terminar su carrera universitaria. Si la

media muestral resultante de las contestaciones fue de 5,2 años con una desviación típica muestral de 1,2 años, obtenga:

(a) Un estimador por intervalo, al 95% de confianza, para el tiempo medio que tardan los estudiantes universitarios en terminar sus carreras.

(b) Un estimador por intervalo similar, al 99% de confianza.

2. El responsable del departamento de transporte de una empresa de paquetería de Nueva York ha recibido quejas acerca de los tiempos que tardan los destinatarios de California en recibir los envíos. Para contrastar la veracidad de las quejas, dicho responsable seleccionó una muestra de 12 órdenes de envío a California e hizo un seguimiento de los mismos para averiguar el tiempo transcurrido hasta su recepción. Los datos resultantes fueron

15, 20, 10, 11, 7, 12, 9, 14, 12, 8, 13, 16

(a) Obtenga un estimador por intervalo, al 90% de confianza, para el tiempo medio que tardan en recibirse los envíos con destino en California.

(b) Calcule un estimador por intervalo similar, al 95% de confianza.

3. Se ha diseñado una encuesta para estimar μ, el salario medio de los ejecutivos de banca de nivel medio. La muestra seleccionada, compuesta por 15 ejecutivos, proporcionó los siguientes salarios anuales (en unidades de 1000 dólares):

88, 121, 75, 39, 52, 102, 95, 78, 69, 82, 80, 84, 72, 115, 106

Obtenga un estimador por intervalo para μ, con un nivel de confianza del:

(a) 90%

(b) 95%

(c) 99%

4. El número de viajeros diarios de una determinada línea de autobús interurbana, en 12 días distintos seleccionados aleatoriamente, fueron

47, 66, 55, 53, 49, 65, 48, 44, 50, 61, 60, 55

(a) Estime el número medio de viajeros diarios de la línea.

(b) Estime la desviación típica del número medio diario de viajeros de la línea citada.

(c) Obtenga un estimador por intervalo, al 95% de confianza, para el número medio de viajeros diarios.

5. Utilice los datos del problema 1 de la sección 8.2 para obtener un estimador por intervalo para el peso medio del total de los participantes en el maratón de Boston de 2004, con un nivel de confianza del:

(a) 95%

(b) 99%

6. En una muestra de 30 transistores de General Electric se observó que la duración media de los mismos había sido de 1210 horas, con una desviación típica de 92 horas. Obtenga un estimador por intervalo para la vida media de la totalidad de los transistores de General Electric, con un nivel de confianza del:

   (a) 90%

   (b) 95%

   (c) 99%

7. En el problema 10 de la sección 8.4 calcule un estimador por intervalo para la temperatura media de inflamación del tipo de sillas allí citado, con un 95% de confianza.

8. Los siguientes datos muestran los puntos obtenidos por el equipo perdedor en 7 partidos de fútbol americano de la Super Copa elegidos aleatoriamente:

   10, 16, 20, 17, 31, 19, 14

   Construya un estimador por intervalo, al 95% de confianza, para la puntuación media de los equipos perdedores sobre el total de los partidos de la Super Copa.

9. Los datos siguientes reflejan los puntos obtenidos por los vencedores de 8 torneos de Maestros de Golf elegidos aleatoriamente:

   285, 279, 280, 288, 279, 286, 284, 279

   Utilice estos datos para construir un estimador por intervalo para la puntuación media de todos los vencedores del Torneo de Maestros, con una confianza del 90%.

10. Todos los estudiantes de una determinada escuela deben realizar una prueba psicológica. Para conocer el tiempo medio que los estudiantes emplean en realizar la prueba se seleccionó una muestra aleatoria de 20 estudiantes. Si el tiempo medio que emplearon en la prueba fue de 12,4 minutos con una desviación típica muestral de 3,3 minutos, obtenga un estimador por intervalo, al 95% de confianza, para la media de los tiempos que tardan los estudiantes de la escuela en realizar la prueba.

11. Una compañía con una gran flota de coches decide asegurarlos contra accidentes. Para determinar el coste medio de reparación por accidente se ha seleccionado una muestra aleatoria de 16 accidentes. Si el coste medio de reparación de estos accidentes fue de 2200 $ con una desviación típica muestral de 800 $, obtenga un estimador por intervalo para el coste medio del total de accidentes, al 90% de confianza.

12. Una antropóloga ha medido las alturas (en pulgadas) de una muestra aleatoria de 64 hombres de una determinada tribu. La media muestral resultó ser 72,4, con una desviación típica muestral de 2,2. Obtenga un estimador por intervalo para la altura media de los hombres de la tribu, con un nivel de confianza del:

   (a) 95%

   (b) 99%

13. Para determinar la duración media de las llamadas telefónicas realizadas al medio día, una compañía de comunicaciones ha seleccionado una muestra aleatoria de 1200 llamadas. La duración media de éstas fue de 4,7 minutos y su desviación típica de 2,2 minutos. Calcule un estimador por intervalo para la duración media del total de las llamadas citadas, con un nivel de confianza del:

    (a) 90%

    (b) 95%

14. Independientemente, 20 estudiantes de Ciencias midieron el punto de licuación del plomo. La media muestral y la desviación típica muestral de dichas medidas fueron 330,2 ºC y 15,4 ºC, respectivamente. Construya un estimador por intervalo para el punto de licuación real del plomo, con un nivel de confianza del:

    (a) 95%

    (b) 99%

15. Una muestra aleatoria de 300 cuentas de titulares de tarjetas VISA en CitiBank mostró que el débito medio muestral era de 1220 $, con una desviación típica muestral de 840 $. Construya un estimador por intervalo para el débito medio de los titulares de tarjeta VISA en CitiBank, al 95% de confianza.

16. Para obtener información acerca del número medio de años de servicio de los policías de Chicago se ha seleccionado una muestra de 46 de ellos. La media de sus tiempos de servicio fue de 14,8 años, con una desviación típica muestral de 8,2 años. Obtenga un estimador por intervalo para el número medio de años de servicio de los policías de Chicago, con un nivel de confianza del:

    (a) 90%

    (b) 95%

    (c) 99%

17. La siguiente sentencia fue mantenida por un "experto" en Estadística: "Si se extrajera una muestra de tamaño 9 de una población normal con media μ, se podría asegurar, con un 95% de certeza, que μ estaría dentro del intervalo $\overline{X} \pm 1{,}96S/3$, siendo $\overline{X}$ la media muestral y S la desviación típica muestral." ¿Esta afirmación es correcta?

18. El modelo geométrico de paseo aleatorio para el precio de una acción o un bien asume que las diferencias sucesivas entre los logaritmos de los precios de cierre diarios constituyen una muestra aleatoria procedente de una distribución normal. Esto implica que la tasa de cambio entre los precios de cierre sucesivos forman una muestra aleatoria procedente de una población determinada (al contrario de lo que ocurría en el modelo lineal de paseo aleatorio visto en el problema 13 de la sección 8.4, donde se suponía que las diferencias diarias constituían una muestra aleatoria). Así, por ejemplo, bajo el modelo geométrico de paseo aleatorio, existe la misma posibilidad de que el precio suba de 100 hasta 102 que de 50 a 51.

    Los siguientes datos muestran los logaritmos y las diferencias sucesivas entre los logaritmos de los precios de cierre del crudo en 20 días consecutivos laborables de

1994. Si se asume que el modelo geométrico de paseo aleatorio es correcto, utilice estos datos para construir un intervalo de confianza para la media poblacional, al 95% de confianza.

Precio	log(precio)	Diferencias entre log(precio)
17,50	2,862201	
17,60	2,867899	5,697966E−03
17,81	2,87976	1,186109E−02
17,67	2,871868	−7,891655E−03
17,53	2,863914	−7,954597E−03
17,39	2,855895	−8,018494E−03
17,12	2,840247	−1,564789E−02
16,71	2,816007	−2,424002E−02
16,70	2,815409	−5,986691E−04
16,83	2,823163	7,754326E−03
17,21	2,84549	2,232742E−02
17,24	2,847232	1,741886E−03
17,22	2,846071	−1,16086E−03
17,67	2,871868	2,579689E−02
17,83	2,880883	9,01413E−03
17,67	2,871868	−9,01413E−03
17,55	2,865054	−6,81448E−03
17,68	2,872434	7,380247E−03
17,98	2,88926	1,682591E−02
18,39	2,911807	2,254701E−02

19. Se han probado doce bombillas, y sus tiempos de funcionamiento (en horas) fueron los siguientes:

    35,6, 39,2, 18,4, 42,0, 45,3, 34,5, 27,9, 24,4, 19,9, 40,1, 37,2, 32,9

    (a) Obtenga un estimador por intervalo para la vida media de las bombillas del mismo tipo que las probadas, al 95% de confianza.

    (b) Se ha mantenido que los resultados de este experimento indican que: “Se puede asegurar, con un 99% de certeza, que la vida media de las bombillas de ese tipo supera las 30 horas”. ¿Es correcta esta sentencia?

20. Los profesores de una escuela pidieron al director que averiguara el número medio por alumno de los días absentismo escolar durante el curso anterior. En lugar de consultar los datos registrados para el total de los alumnos, el director seleccionó una muestra aleatoria de 50 de ellos, con los que obtuvo que el número medio de días de absentismo escolar durante el curso anterior había sido 8,4, con una desviación típica muestral de 5,1.

(a) Con estos datos, al 95% de confianza, ¿qué estimador por intervalo se obtendría para el número medio de días de ausencia sobre el total de los estudiantes de la escuela?

(b) En una posterior reunión de profesores el director mantuvo: "Puedo asegurar, con un 95% de confianza, que el número medio de días de absentismo escolar por alumno en el curso anterior fue menor que _____." Rellene el valor que falta.

21. En el problema 3 supongamos que se desea asegurar, con un 99% de confianza, que el salario medio es mayor que ν_1. ¿Cuál es el valor apropiado de ν_1? ¿Cuál sería el valor de ν_2 si se quisiera mantener, con una confianza del 99%, que el salario medio es menor que ν_2?

22. En el problema 2 obtenga un valor que supere, con un 95% de confianza, el tiempo medio que tardan los destinatarios de California en recibir un envío.

23. Para convencer a un comprador potencial sobre la viabilidad de cierta empresa, un ejecutivo ha diseñado un muestreo de las ventas diarias. Sobre una muestra de 14 días se observaron los siguientes valores de ventas (en unidades de 100 dólares):

33, 12, 48, 40, 26, 17, 29, 38, 34, 41, 25, 51, 49, 34

Si el ejecutivo pretende presentar estos datos de la forma más favorable posible, ¿debería mostrar un intervalo de confianza o una cota unilateral de confianza? Si fuera esta última, ¿debería ser una cota superior o inferior? La siguiente frase: "Puedo mantener, con un 95% de confianza, que…", ¿cómo debería completarla?

24. Para calmar las reticencias de un grupo de ciudadanos preocupados por la polución atmosférica de su barrio, un inspector medioambiental obtuvo los datos de una muestra aleatoria de concentraciones de monóxido de carbono. Estos datos, en partes por millón, fueron los siguientes:

101,4, 103,3, 101,6, 111,6, 98,4, 95,0, 93,6

Si estos datos al inspector le parecen razonablemente bajos, hablando "con un 99% de confianza", ¿cómo debería presentarlos a dicho grupo de ciudadanos?

8.7 Estimadores por intervalo de una proporción poblacional

Supongamos que pretendemos obtener un estimador por intervalo para p, la proporción de individuos que en una población de gran tamaño presentan una determinada característica. Supongamos, además, que se ha seleccionado una muestra aleatoria de tamaño n, en la que se observó que X de sus miembros presentaban la característica. Si la proporción de elementos muestrales que presentan la característica se denota por $\hat{p} = X/n$, se tendrá que, tal como se vio en la sección 8.3, el valor esperado y la desviación típica de $\hat{p}$ son

$$E[\hat{p}] = p$$

$$\mathrm{SD}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

Si n es lo suficientemente grande, de forma que tanto np como $n(1 - p)$ son mayores que 5, se puede utilizar la aproximación de la distribución normal a la binomial para obtener un estimador por intervalo, al $100(1 - \alpha)\%$ de confianza, para p, que vendrá dado por

$$\hat{p} \pm z_{\alpha/2} \text{ SD } (\hat{p})$$

Aunque la desviación típica de $\hat{p}$ no se conoce, puesto que en la expresión de SD $(\hat{p})$ aparece la proporción desconocida p, se puede estimar si se reemplaza p por su estimador $\hat{p}$. Esto es, se puede estimar SD $(\hat{p})$ mediante $\sqrt{\hat{p}(1 - \hat{p})/n}$. Esto permite escribir lo siguiente:

Un estimador por intervalo aproximado para p, al $100(1 - \alpha)\%$ de confianza, viene dado por

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

donde $\hat{p}$ es la proporción de miembros de la muestra de tamaño n que presentan la característica.

Ejemplo 8.15 En una muestra aleatoria de 100 estudiantes de una determinada universidad, 82 de ellos manifestaron que no eran fumadores. Sobre esta base obtenga un estimador por intervalo para p, la proporción de estudiantes no fumadores de la universidad, al 99% de confianza.

Solución Puesto que $100(1 - \alpha)\% = 0{,}99$ si $\alpha = 0{,}01$, se necesita obtener el valor de $z_{\alpha/2} = z_{0{,}005}$, el cual es igual a 2,576, como puede verse en la tabla D.2. Así pues, el estimador por intervalo para p, al 99% de confianza, es

$$0{,}82 \pm 2{,}576\sqrt{\frac{0{,}82(1 - 0{,}82)}{100}}$$

es decir,

$$0{,}82 \pm 0{,}099$$

Así pues, se puede mantener, con un 99% de confianza, que el porcentaje de no fumadores está comprendido entre 72,1 y 91,9%. ■

Ejemplo 8.16 El 24 de diciembre de 1991, el periódico *New York Times* publicó una encuesta, de la que se concluía que el 46% de la población estaba a favor de la forma en que el presidente Bush estaba llevando la economía de Estados Unidos, con un margen de error de $\pm 3\%$. ¿Qué significa esto? ¿Se puede saber el número de personas encuestadas?

Solución Es una práctica común en los medios de comunicación presentar estimadores por intervalo con un 95% de confianza, a menos que explícitamente se mencione otro nivel de confianza. Puesto que $z_{0,025} = 1,96$, un estimador por intervalo para p, al 95% de confianza, viene dado por

$$\hat{p} \pm 1,96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

donde n es el tamaño muestral. Dado que $\hat{p}$, la proporción de elementos de la muestra que se declararon a favor de la forma en que el presidente Bush estaba manejando la economía, era igual a 0,46, se tiene que el estimador por intervalo, con una confianza del 95%, para p, la proporción poblacional a favor de Bush en la fecha citada, es

$$0,46 \pm 1,96\sqrt{\frac{(0,46)(0,54)}{n}}$$

Puesto que el margen de error se dice que fue de $\pm 3\%$, se tiene que

$$1,96\sqrt{\frac{(0,46)(0,54)}{n}} = 0,03$$

Si se elevan al cuadrado los dos miembros de esta igualdad se obtiene que

$$(1,96)^2\frac{(0,46)(0,54)}{n} = (0,03)^2$$

o, equivalentemente,

$$n = \frac{(1,96)^2(0,46)(0,54)}{(0,03)^2} = 1060,3$$

Es decir, se encuestó aproximadamente a 1060 personas, entre las que el 46% se mostraron a favor de la forma en que el presidente Bush llevaba la economía de Estados Unidos. ■

8.7.1 Longitud del intervalo de confianza

Puesto que el intervalo para p, al $100(1-\alpha)\%$ de confianza, tiene por extremos

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \qquad \text{y} \qquad \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

la longitud del intervalo es la que se indica a continuación.

La Estadística en perspectiva

Estudio de caso

El programa de ayuda a familias con hijos dependientes reconoce que los errores son inevitables y que, por consiguiente, no todas las familias a las que subvenciona reúnen los requisitos exigidos. Pese a ello, el Estado de California mantiene que los municipios tienen la responsabilidad de controlar que los requisitos se cumplan, y ha fijado una tasa de error máxima de un 4%. Esto significa que si, en un determinado municipio, se descubre que el porcentaje de familias que no reúnen los requisitos supera el 4%, se penaliza al municipio en una cuantía que depende del porcentaje de error. Dado que el Estado de California no tiene recursos para comprobar si cada familia subvencionada reúne los requisitos exigidos, utiliza un muestreo aleatorio para estimar los porcentajes de error.

En 1981, se seleccionó una muestra de familias subvencionadas del municipio de Alameda y se encontró que 9 de ellas no cumplían las condiciones exigidas. Basándose en esto, se estimó que el porcentaje de familias subvencionadas que no cumplían los requisitos era del $100 \times 9/152 = 5{,}9\%$, por lo que se impuso una penalización de 949 597 $ al municipio citado. Éste apeló la decisión ante el juzgado, y razonó que 9 familias erróneamente subvencionadas en una muestra de 152 de ellas no proporcionaban una suficiente evidencia de que el porcentaje de error, sobre el total de familias, superara el 4%. Con el dictamen técnico de un equipo de expertos en Estadística, el juez decidió que no era justo utilizar el porcentaje 5,9 del estimador puntual como el porcentaje de error verdadero en el municipio. El juez decidió que sería más justo utilizar el extremo inferior del estimador por intervalo, con un 95% de confianza. Puesto que, con esta confianza, el estimador por intervalo para el porcentaje de familias subvencionadas que no reunían los requisitos exigidos es

$$0{,}059 \pm 1{,}96\sqrt{\frac{0{,}059(1 - 0{,}059)}{152}} = 0{,}059 \pm 0{,}037$$

se tiene que el extremo inferior del intervalo es $0{,}059 - 0{,}037 = 0{,}022$. Así pues, dado que este valor es menor que 0,04, el juez anuló la sanción impuesta al dictaminar que era una penalización indebida.

La longitud del intervalo de confianza, al $100(1 - \alpha)\%$ de confianza, es

$$2z_{\alpha/2}\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

siendo $\hat{p}$ la proporción de elementos de la muestra que presentan la característica.

Dado que se puede comprobar que el producto $\hat{p}(1-\hat{p})$ es siempre menor que 1/4, se desprende de la expresión anterior que $2z_{\alpha/2}\sqrt{1/(4n)}$ es una cota superior para la longitud del intervalo de confianza, lo cual es equivalente a

$$\text{Longitud del intervalo al } 100(1-\alpha)\% \text{ de confianza} \leq \frac{z_{\alpha/2}}{\sqrt{n}}$$

Se puede utilizar la cota anterior para calcular el tamaño de muestra que se necesita para obtener un intervalo de confianza cuya longitud sea menor que un determinado valor prefijado de antemano. Por ejemplo, supongamos que se desea determinar un tamaño muestral que garantice que la longitud del intervalo, al $100(1-\alpha)\%$ de confianza, sea menor que un valor b. A partir de la desigualdad anterior se puede concluir que cualquier tamaño muestral n que verifique que

$$\frac{z_{\alpha/2}}{\sqrt{n}} < b$$

permite garantizar que la longitud del intervalo de confianza es menor que b. Es decir, se debe elegir n de forma que

$$\sqrt{n} > \frac{z_{\alpha/2}}{b}$$

Así pues, si se elevan al cuadrado los dos miembros de la desigualdad anterior, se tiene que n debe cumplir que

$$n > \left(\frac{z_{\alpha/2}}{b}\right)^2$$

Ejemplo 8.17 ¿Qué tamaño muestral se necesita para garantizar que la longitud del estimador por intervalo, al 90% de confianza, para p sea menor que 0,01?

Solución Para poder asegurar que la longitud del estimador por intervalo para p, al 90% de confianza, es menor que 0,01, se necesita elegir n de forma que

$$n > \left(\frac{z_{0,05}}{0,01}\right)^2$$

Dado que $z_{0,05} = 1,645$, se obtiene que

$$n > (164,5)^2 = 27\,062,25$$

Esto es, para que se pueda asegurar que el estimador por intervalo, al 90% de confianza, tenga una longitud menor que 0,01, el tamaño muestral debe superar el valor 27,063.

Si se denota como L a la longitud del intervalo de confianza para p,

$$\underset{\hat{p}-\frac{L}{2}}{\leftarrow} \qquad \underset{\hat{p}}{L} \qquad \underset{\hat{p}+\frac{L}{2}}{\rightarrow}$$

está claro que cualquier punto del intervalo dista de $\hat{p}$ como máximo $L/2$, dado que el punto medio del intervalo es $\hat{p}$. Así pues, en el ejemplo 8.17, se podrá asegurar que, con una muestra de tamaño superior a 27,063, la proporción muestral observada diferirá de la proporción poblacional real menos de 0,005, con un 90% de confianza. ■

8.7.2 Cotas inferior y superior de confianza

Fácilmente se pueden obtener las cotas inferior y superior de confianza para p que se indican a continuación.

Una cota inferior para p, al $100(1-\alpha)\%$ de confianza, viene dada por

$$\hat{p} - z_\alpha \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Esto es, se puede mantener, con un $100(1-\alpha)\%$ de confianza, que la proporción de elementos de la población que presentan la característica es mayor que el valor de la cota anterior.

De igual forma, una cota inferior para p, con un $100(1-\alpha)\%$ de confianza, viene dada por

$$\hat{p} + z_\alpha \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Es decir, con un $100(1-\alpha)\%$ de confianza, se puede asegurar que la proporción de elementos de la población que presentan la característica es menor que el valor de la anterior cota.

Ejemplo 8.18 En una muestra de 125 trabajadores de una gran ciudad, 42 de ellos manifestaron no estar satisfechos con sus condiciones de trabajo. Obtenga una cota inferior para la proporción poblacional de trabajadores insatisfechos con sus condiciones de trabajo, al 95% de confianza.

Solución Puesto que $z_{0,05} = 1,645$ y $42/125 = 0,336$, la cota inferior pedida viene dada por

$$0,336 - 1,645 \sqrt{\frac{0,336(0,664)}{125}} = 0,2665$$

Es decir, se puede asegurar, con un 95% de confianza, que más del 26,6% de la población de trabajadores de la ciudad no está satisfecho con sus condiciones de trabajo. ■

Problemas

1. En una muestra de 500 personas de California, 302 de ellas se mostraron a favor de la pena de muerte. Obtenga un estimador por intervalo, al 99% de confianza, para la proporción poblacional de individuos a favor de la pena de muerte.
2. Se sabe que las personas que sufren un ataque de corazón por primera vez tienen una mayor propensión que el resto a sufrir nuevos ataques de corazón en el plazo de un año. Para estimar la proporción de ellas que sufren nuevos ataques de corazón en el plazo de

un año se seleccionó una muestra de 300 personas que sufrieron un ataque recientemente y se les hizo un seguimiento durante un año.

(a) Si 46 de ellas sufrieron un nuevo ataque, obtenga un estimador por intervalo, al 95% de confianza, para la proporción poblacional indicada.

(b) Repita el apartado (a) y asuma que 92 individuos de la muestra volvieron a sufrir un nuevo ataque durante el año de seguimiento.

3. Para estimar p, la proporción poblacional de bebés recién nacidos que son varones, se anotó el sexo de cada uno de los bebés de una muestra aleatoria de 10 000 recién nacidos. Si 5106 de ellos fueron varones, obtenga un estimador por intervalo para p, con un nivel de confianza del:

   (a) 90%

   (b) 99%

4. En 1980, una encuesta sobre 1200 votantes dio a Ronald Reagan un 57% de los votos. Obtenga un estimador por intervalo, al 99% de confianza, para la proporción poblacional de votantes a favor de Reagan en el momento de la encuesta.

5. En una muestra de 100 residentes de Los Ángeles, 64 de ellos se mostraron a favor de una legislación estricta de control de armas. Determine, con un 95% de confianza, un estimador por intervalo para la proporción poblacional de residentes de Los Ángeles a favor del control de armas.

6. En una muestra de 100 recientes doctores en Ciencias, 42 de ellos se mostraron optimistas acerca de su futuro investigador. Encuentre un estimador por intervalo para la proporción poblacional de recientes doctores que son optimistas, con un nivel de confianza del:

   (a) 90%

   (b) 99%

7. En el problema 1 de la sección 8.3, calcule un estimador por intervalo para la proporción de norteamericanos que en 1981 creían que el partido comunista ganaría unas elecciones libres en la Unión Soviética, con una confianza del 95%.

8. Con los datos del problema 4 de la sección 8.3 obtenga un estimador por intervalo, al 90% de confianza, para la probabilidad de ganar al solitario.

9. Un importador de vinos tiene la oportunidad de comprar un gran remesa de botellas de vino Chateau Lafite Rothschild, de la cosecha de 1947. Debido a la edad del vino algunas de las botellas pueden haberse avinagrado. Sin embargo, la única forma de averiguar si una botella está en buen estado es abrirla y probarla. Por este motivo, el importador ha acordado con el vendedor seleccionar aleatoriamente 20 botellas y abrirlas. Supongamos que en 3 de ellas el vino estaba estropeado. Calcule un estimador por intervalo para la proporción de botellas de la remesa que no están en buen estado, al 95% de confianza.

10. Se recogió una muestra de 100 tazas de café servidas por una máquina automática y se midieron las cantidades de café de cada taza. Supongamos que en 9 de ellas el contenido servido era menor que lo anunciado en la máquina. Con un 90% de confianza, construya un estimador por intervalo para la proporción de tazas, sobre el total de las suministradas por la máquina, cuyo contenido es menor que el especificado.

11. En una muestra de 400 libreros, 335 de ellos eran mujeres. Obtenga un estimador por intervalo, al 95% de confianza, para la proporción poblacional de libreros que son mujeres.

12. Una muestra aleatoria de 300 escritores incluía 117 hombres. Calcule un estimador por intervalo para la proporción de hombres, sobre el total de escritores.

13. Una muestra aleatoria de 9 Estados de Estados Unidos (Virginia Occidental, Nueva York, Idaho, Tejas, Nuevo México, Indiana, UTA, Maryland y Maine) mostró que en 2 de ellos la renta per cápita en 1990 superaba 20 000 $. Construya un estimador por intervalo, al 90% de confianza, para la proporción de Estados de Estados Unidos cuya renta per cápita en 1990 excedía los 20 000 $.

14. Una muestra aleatoria de 1000 psicólogos incluía a 457 hombres. Calcule un estimador por intervalo, al 95% de confianza, para la proporción de varones sobre el total de los psicólogos.

15. En una muestra aleatoria de 500 contables existían 42 de origen africano, 18 hispanos y 246 mujeres. Sobre el total de los contables obtenga un estimador por intervalo, al 95% de confianza, para la proporción de ellos que son:

    (a) de origen africano

    (b) hispanos

    (c) mujeres

16. En una encuesta llevada a cabo el 22 de enero de 2004, sobre las 600 personas muestreadas, 450 de ellas se mostraron a favor de la guerra de Irak. Obtenga un estimador por intervalo para p, la proporción poblacional que personas que estaban a favor de la guerra de Irak en la fecha indicada, con una confianza del:

    (a) 90%

    (b) 95%

    (c) 99%

17. La encuesta del problema 16 se publicó el 28 de enero de 2004 en el periódico *Crónica de San Francisco*; donde se mantuvo que: "El 75% de la población está a favor de la guerra de Irak con un margen de error de ± 4 puntos porcentuales."

    (a) Explique por qué el periódico debería haber indicado que el margen de error era de $\pm 3{,}46$ puntos.

    (b) Explique en qué se equivocó el periódico para llegar al valor de ± 4 puntos.

18. Una reciente encuesta llevada a cabo por un periódico mostró que el candidato A vencía al candidato B por 53 a 47 puntos porcentuales, con un margen de error de ± 4 puntos. El periódico aseguró a sus lectores que, dado que los 6 puntos de separación entre ellos era mayor que el margen de error, el candidato A ganaría con seguridad. ¿Es correcto este razonamiento?

19. Una compañía de investigación de mercado está interesada en conocer la proporción de hogares en los que se verá un determinado acontecimiento deportivo. Para conseguirlo, planea hacer una encuesta por teléfono a una muestra de hogares seleccionados aleatoriamente.

(a) Si la compañía desea que su estimador tenga un margen de error de $\pm 0{,}02$, ¿qué tamaño de muestra se necesita?

(b) Supongamos que se selecciona una muestra de tamaño acorde con el resultado obtenido en el apartado (a). Si el 23% de la muestra vio el acontecimiento deportivo, ¿se puede intuir cuál será la longitud de dicho intervalo, con un 90% de confianza: exactamente igual, mayor o menor que 0,02?

(c) Obtenga el intervalo de confianza del apartado (b).

20. ¿Cuál es el mínimo número de certificados de muerte que se deben muestrear aleatoriamente para estimar la proporción de la población de Estados Unidos que muere de cáncer si se desea que, con un 95% de confianza, el valor estimado difiera del real en 0,01, como máximo?

21. Supongamos que se sabe que, en el problema 20, aproximadamente un 20% de las muertes se deben al cáncer. Con esta información determine aproximadamente el número de certificados de muerte que se han de muestrear para poder asegurar que se cumple lo requerido en el problema 20.

22. Utilice los datos del problema 14 para obtener una cota inferior, con una confianza del 95%, para la proporción de varones sobre el total de psicólogos existentes.

23. Utilice los datos del problema 11 para obtener una cota superior, con una confianza del 95%, para la proporción de mujeres sobre el total de libreros existentes.

24. Un productor está pensando poner un anuncio donde se mantiene que más de un x por ciento de los usuarios de sus productos está satisfecho con ellos. Para determinar x se entrevistó a una muestra aleatoria de 500 usuarios. Si el 92% de ellos mostraron su satisfacción con el producto y el productor desea que el anuncio tenga una confianza del 95%, ¿qué valor de x debería figurar en el anuncio? ¿Y si se quisiera que la confianza en la veracidad del mismo fuera del 90%?

25. Utilice los datos del problema 15 para obtener una

(a) cota inferior, al 90% de confianza,

(b) cota superior, al 90% de confianza,

para p, la proporción de afroamericanos o hispanos que existen, sobre el total de contables de Estados Unidos.

26. En el problema 16, obtenga una

(a) cota superior, al 95% de confianza,

(b) cota inferior, al 95% de confianza,

para p, la proporción de la población que estaba a favor de la guerra de Irak en el momento de la encuesta.

27. Supongamos que, en el problema 9, el importador considera que obtendrá beneficios con la compra de la remesa del vino, siempre que el porcentaje de botellas mal conservadas sea menor del 20%. Con los datos del problema indicado, ¿puede el importador obtener beneficios en la compra? Asuma una confianza del

(a) 95%

(b) 99%

28. Teniendo en cuenta los datos del problema 3 rellene los valores que faltan en las siguientes frases:

    (a) Con un 95% de confianza, más de un _____ por ciento de los residentes de Los Ángeles está a favor del control de armas.

    (b) Con un 95% de confianza, menos de un _____ por ciento de los residentes de Los Ángeles está a favor del control de armas.

Términos clave

Estimador: Un estadístico utilizado para aproximar un parámetro de la población. También se le denomina *estimador puntual*.

Valor estimado: Valor observado del estimador.

Estimador insesgado: Estimador cuya esperanza coincide con el parámetro que se desea estimar.

Error estándar de un estimador (insesgado): Desviación típica del estimador. Es un indicador de la diferencia que se puede esperar que exista entre el estimador y el parámetro que se desea estimar.

Estimador por intervalo de confianza: Intervalo cuyos extremos se determinan a partir de los datos muestrales. El parámetro está contenido en el intervalo con un grado determinado de confianza. Por lo general, el punto medio del intervalo coincide con el estimador puntual del parámetro.

Nivel de confianza del $100(1 - \alpha)\%$: La proporción de veces que, a largo plazo, el intervalo contiene el parámetro que se va a estimar. Equivalentemente, antes de observar los datos, el estimador por intervalo contendrá el parámetro con una probabilidad de $1 - \alpha$; tras haber observado los datos, el intervalo estimado resultante contiene al parámetro con un $100(1 - \alpha)\%$ de confianza.

Cota inferior de confianza: Valor determinado a partir de los datos muestrales para el que se puede mantener que es menor que el parámetro, con un determinado grado de confianza.

Cota superior de confianza: Valor determinado a partir de los datos muestrales, para el que se puede mantener que supera al parámetro, con un determinado grado de confianza.

Variable aleatoria *t*: Si $X_1, \ldots, X_n$ es una muestra procedente de una población con media μ, se dice que la variable aleatoria

$$\sqrt{n}\,\frac{\overline{X} - \mu}{S}$$

es una variable aleatoria t con $n - 1$ grados de libertad, donde $\overline{X}$ y S representan la media muestral y la desviación típica muestral, respectivamente.

Resumen

La media muestral $\overline{X}$ es un estimador insesgado de la media poblacional μ. Su desviación típica, también conocida como *error estándar* de $\overline{X}$ como estimador de μ, viene dada por

$$\mathrm{SD}(\overline{X}) = \frac{\sigma}{\sqrt{n}}$$

siendo σ la desviación típica de la población.

El estadístico $\hat{p}$, igual a la proporción de elementos de la muestra que poseen una determinada característica, es un estimador de p, la proporción de elementos de la población que presentan la característica citada. El error estándar de este estimador es

$$\mathrm{SD}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

siendo n el tamaño muestral. Este error estándar se puede calcular mediante

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

La varianza muestral S^2 es un estimador de la varianza poblacional σ^2. En concordancia, se utiliza la desviación típica muestral para estimar la desviación típica poblacional σ.

Si $X_1, \ldots, X_n$ es una muestra procedente de una población normal con desviación típica conocida σ,

$$\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

es un estimador por intervalo para la media poblacional μ, al $100(1-\alpha)\%$ de confianza. La longitud de este intervalo, es decir,

$$2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

será menor o igual que b si el tamaño muestral n verifica que

$$n \geq \left(\frac{2z_{\alpha/2}\sigma}{b}\right)^2$$

Una cota inferior para μ, al $100(1-\alpha)\%$ de confianza, viene dada por

$$\overline{X} - z_{\alpha}\frac{\sigma}{\sqrt{n}}$$

Esto es, se puede asegurar, con un $100(1-\alpha)\%$ de confianza, que

$$\mu > \overline{X} - z_{\alpha}\frac{\sigma}{\sqrt{n}}$$

Una cota superior para μ, al $100(1 - \alpha)\%$ de confianza, es

$$\overline{X} + z\alpha_{\alpha} \frac{\sigma}{\sqrt{n}}$$

Es decir, se puede asegurar, con un $100(1 - \alpha)\%$ de confianza, que

$$\mu < \overline{X} + z_{a} \frac{\sigma}{\sqrt{n}}$$

Si $X_1, \ldots, X_n$ es una muestra procedente de una población normal con desviación típica desconocida, un estimador por intervalo para μ, al $100(1 - \alpha)\%$ de confianza, es

$$\overline{X} \pm t_{n-1,\,\alpha/2} \frac{S}{\sqrt{n}}$$

En el intervalo anterior, el valor $t_{n-1,\,\alpha/2}$ es tal que

$$P\{T_{n-1} > t_{n-1,\,\alpha/2}\} = \frac{\alpha}{2}$$

siendo T_{n-1} una variable aleatoria t con $n - 1$ grados de libertad.

Las cotas inferior y superior para μ, al $100(1 - \alpha)\%$ de confianza, vienen dadas, respectivamente, por

$$\overline{X} - t_{n-1,\,\alpha} \frac{S}{\sqrt{n}}$$

y

$$\overline{X} + t_{n-1,\,\alpha} \frac{S}{\sqrt{n}}$$

Para obtener un estimador por intervalo de confianza para p, la proporción de elementos de una población grande que presentan una determinada característica, se debe seleccionar una muestra de tamaño n. Si $\hat{p}$ representa la proporción de elementos de la muestra que presentan la característica, un estimador por intervalo aproximado para p, al $100(1 - \alpha)\%$ de confianza, es

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

La longitud de este intervalo verifica con seguridad que

$$\text{Longitud del intervalo de confianza} \leq \frac{z_{\alpha/2}}{\sqrt{n}}$$

La distancia desde el centro a los extremos del intervalo, al 95% de confianza, es decir, el valor de $1{,}96\sqrt{\hat{p}(1 - \hat{p})/n}$, se denomina habitualmente *margen de error* del intervalo. Por ejemplo, supongamos que un periódico publica que, en una nueva encuesta, el 64% de la población se consideran a sí mismos conservadores, con un margen de error de $\pm 3\%$. Esto significa que la encuesta permite obtener un estimador por intervalo, para la proporción

poblacional de personas que se consideran conservadoras, igual a 0,64 ± 0,03, con una confianza del 95%.

Problemas de repaso

1. ¿En qué caso se obtendrá un estimador más preciso para μ?
   (a) Con una muestra de tamaño n extraída de una población con media 2μ y varianza σ^2.
   (b) Con una muestra de tamaño $2n$ extraída de una población con media μ y desviación típica σ.
2. Las longitudes de un tipo de cojinetes de bolas se distribuyen normalmente con una desviación típica de 0,5 milímetros.
   (a) ¿Qué tamaño de muestra se necesita si se pretende que el estimador de la longitud media poblacional de los cojinetes difiera de la longitud media muestral en ±0,1 milímetros?
   (b) Repita (a) si se desea que el estimador difiera de la longitud media muestral en ±0,01 milímetros.
   (c) Si los valores de una muestra de tamaño 8 son

   4,1, 4,6, 3,9, 3,3, 4,0, 3,5, 3,9, 4,2

   obtenga un estimador por intervalo para la longitud media muestral, al 95% de confianza.
3. Se pidió a los miembros de una muestra aleatoria de 50 personas que registraran el tiempo que habían empleado viendo televisión durante una semana. Si la media muestral y la desviación típica muestral de los datos resultantes fueron 24,4 horas y 7,4 horas, respectivamente, obtenga un estimador por intervalo, al 95% de confianza, para el tiempo medio empleado en ver la televisión durante la semana citada sobre el total de individuos de la población.
4. Utilice los 30 primeros valores de datos del Apéndice A para obtener un estimador por intervalo, al 90% de confianza, para el nivel medio de colesterol de todos los estudiantes de la lista. Ahora, divida los 30 valores de datos en dos grupos, uno de las mujeres y otro de los hombres. Utilice estos datos para obtener separadamente para ambos sexos un estimador por intervalo, al 90% de confianza, para los niveles poblacionales de colesterol en sangre. ¿Qué confianza se puede dar al hecho de que los niveles medios tanto para las mujeres como para los hombres están contenidos en sus respectivos intervalos, al 90% de confianza?
5. Todos los estudiantes del sexto año de primaria del Estado de Washington deben someterse a un test estándar. Para obtener la puntación media de los estudiantes de su distrito, un supervisor escolar selecciona una muestra aleatoria de 100 estudiantes. Si la media muestral de las calificaciones es de 320 puntos y la desviación típica muestral es de 16 puntos, obtenga un estimador por intervalo, al 95% de confianza, para la puntuación media de todos los estudiantes del distrito.

6. Una línea aérea quiere averiguar la proporción de sus usuarios que vuelan por razones de negocios. Si la línea aérea pretende que su estimador difiera del valor correcto en menos de 2 puntos porcentuales, con una confianza del 90%, ¿cuál debe ser el tamaño de la muestra empleada?

7. Los siguientes datos representan el número de consumiciones diarias que se obtienen de una máquina automática en una muestra de 20 días distintos:

   56, 44, 53, 40, 65, 39, 36, 41, 47, 55, 51, 50, 72, 45, 69, 38, 40, 51, 47, 53

   (a) Calcule un estimador por intervalo, al 95% de confianza, para el número medio de consumiciones obtenidas diariamente de la máquina automática.

   (b) Repita el apartado (a), con una confianza del 90%.

8. Se piensa que el periodo en el que se duerme más profundamente, que parece coincidir con el periodo en que los sueños ocurren con mayor frecuencia, se caracteriza por un rápido movimiento de los ojos de la persona que está durmiendo. A un voluntario se le midieron los tiempos del citado movimiento rápido de los ojos en 7 noches distintas, y se obtuvieron los siguientes datos en minutos:

   37, 42, 51, 39, 44, 48, 29

   Obtenga un estimador por intervalo para las duraciones del movimiento rápido de ojos del voluntario en cuestión, al 99% de confianza.

9. Una gran compañía desea analizar su actual política de ayuda sanitaria. En particular está interesada en averiguar el coste medio por maternidad. Supongamos que el coste medio reclamado en una muestra de 24 nacimientos fue de 1840 dólares, con una desviación típica muestral de 740 dólares. Obtenga un estimador por intervalo, al 95% de confianza, para el coste medio por maternidad para el total de los nacimientos subvencionados.

10. Una encuesta que pidió un juez mostró que, en una muestra aleatoria de 300 trabajadores agrarios, 144 de ellos se mostraron a favor de su sindicación. Obtenga un estimador por intervalo, al 90% de confianza, para la proporción de trabajadores agrarios, sobre el total de éstos, que estaban a favor de su sindicación.

11. Se midieron las velocidades, en millas por hora, de nueve bolas rápidas lanzadas por un determinado jugador de béisbol, y se obtuvieron los siguientes datos:

    94, 87, 80, 91, 85, 102, 85, 80, 93

    (a) ¿Cuál es el estimador puntual de la velocidad media de los lanzamientos rápidos de dicho jugador?

    (b) Obtenga un estimador por intervalo, al 95% de confianza, para la velocidad media de los lanzamientos citados.

12. Con una muestra de tamaño 9 se obtuvo una media muestral de 35. Obtenga un estimador por intervalo, al 95% de confianza, para la media poblacional si se sabe que la desviación típica de la población es igual a:

    (a) 3

(b) 6

(c) 12

13. Repita el problema 12 para un tamaño muestral de 36.

14. Los siguientes datos presentan las calificaciones obtenidas en un test por los 18 estudiantes de una muestra aleatoria extraída en cierta universidad.

130, 122, 119, 142, 136, 127, 120, 152, 141,
132, 127, 118, 150, 141, 133, 137, 129, 142

(a) Obtenga un estimador por intervalo para la calificación media del total de los estudiantes de la universidad, con un 90% de confianza.

(b) Construya un estimador por intervalo similar al anterior, al 95% de confianza.

(c) Repita el apartado (b), con una confianza del 99%.

15. Para hacer cumplir la legislación vigente, un gestor escolar necesita calcular la proporción de profesores de educación secundaria que son mujeres. Si en una muestra aleatoria de 1000 profesores, 518 de ellos son mujeres, obtenga un estimador por intervalo, al 95% de confianza.

16. En el problema 15, supongamos que el gestor escolar quisiera obtener un estimador por intervalo, al 99% de confianza, cuya longitud sea, como máximo, 0,03. ¿Qué tamaño muestral sería necesario?

17. La Oficina de Censos de Estados Unidos utiliza una muestra aleatoria de tamaño 50 000 para determinar la tasa nacional de desempleo. ¿Cuál es el mayor margen de error posible?

18. Un investigador desea conocer la proporción de personas que está a favor de un determinado candidato. Si se extrae una muestra aleatoria de tamaño 1600, ¿cuál es el mayor margen de error posible?

19. Un problema interesante en béisbol es analizar si sacrificar una parada de bola constituye una buena estrategia cuando hay un jugador en la primera base y ninguno fuera. Si se asume que quien para la bola saldrá fuera pero que, aún así, podría ser ventajoso para que avance el corredor de la base, se podría comparar la probabilidad de conseguir una carrera con un jugador en la primera base y ninguno fuera, con la probabilidad de conseguir una carrera con un jugador en la segunda base y uno fuera. Los siguientes datos se han extraído de una muestra de partidos importantes de la liga de béisbol de Estados Unidos en 1959 y 1960.

Base ocupada	Número de jugadores fuera	Proporción de casos en los que no se consiguieron carreras	Número total de casos
Primera	0	0,604	1728
Segunda	1	0,610	657

(a) Obtenga un estimador por intervalo, al 95% de confianza, para la probabilidad de puntuar al menos una carrera cuando hay un jugador en la primera base y no existen jugadores fuera.

(b) Obtenga un estimador por intervalo, al 95% de confianza, para la probabilidad de puntuar al menos una carrera cuando hay un jugador en la segunda base y existe un jugador fuera.

20. Utilice los datos del problema 15 para obtener una cota superior para la proporción de mujeres sobre el total de profesores de secundaria, con una confianza del:

(a) 90%

(b) 95%

(c) 99%

21. Repita el problema 20, cuando se desea obtener cotas inferiores similares, con las confianzas citadas. Si uno fuera partidario de resaltar la mayor contratación de mujeres como profesores, ¿qué cota de confianza haría pública: la superior o la inferior?

22. Supongamos que se ha extraído una muestra aleatoria de nueve casas vendidas recientemente en un determinado barrio. Si el precio medio de venta muestral fue de 222 000 $ con una desviación típica muestral de 12 000 $, obtenga una cota superior para el precio medio de venta de las casas del barrio, con un 95% de confianza.

9

Contraste de hipótesis estadísticas

La gran tragedia de la ciencia: la destrucción de una bella hipótesis por un antiestético conjunto de datos.

Thomas H. Huxley, biólogo inglés (*Biogénesis* y *Abiogénesis*)

Todos aprendemos de la experiencia, y la lección en esta ocasión es que nunca se debe perder de vista la alternativa.

Sherlock Holmes, en *Las Aventuras de Black Peter* por Sir Arthur Conan Doyle

En este capítulo se explicará qué se entiende por una hipótesis estadística y se verá cómo se puede contrastar a partir de un conjunto de datos. Se distinguirá entre la hipótesis nula y la hipótesis alternativa. Se explicará qué significa rechazar la hipótesis nula y no rechazarla. Se introducirá el concepto de p valor resultante de un contraste.

A continuación se estudiarán los contrastes relativos a la media de una población normal, tanto cuando la varianza poblacional es conocida como cuando no lo es. Se considerarán los contrastes unilaterales y los bilaterales. Finalmente se presentarán los contrastes relativos a una proporión poblacional.

9.1 Introducción

En los últimos años ha existido una gran controversia sobre los peligros de vivir cerca de un campo electromagnético de alto nivel. Una investigadora, tras escuchar un gran número

de historias anecdóticas que hacían referencia a que, en las comunidades próximas a los citados campos, se detectaban fuertes aumentos de los casos de cáncer, especialmente entre los niños, decidió analizar los posibles peligros. Para ello, primero estudió los mapas de situación de las líneas eléctricas de alta tensión y seleccionó un gran barrio que estaba situado en un área sometida a un campo electromagnético de alto nivel. Después realizó entrevistas en las escuelas, los hospitales y los servicios de sanidad pública de la localidad para averiguar el número de niños que habían sido afectados por el cáncer (de cualquier tipo) en los tres años anteriores. Y llegó a contabilizar que se habían producido 32 casos.

Posteriormente, consultó los datos oficiales de sanidad para averiguar el número de casos de cáncer en niños que se podía esperar que ocurrieran en un barrio de tamaño similar al seleccionado. Así obtuvo que el número medio de casos de cáncer infantil registrados en barrios similares, a lo largo de tres años, era de 16,2 con una desviación típica de 4,7.

Si se compara con el número medio 16,2, ¿los 32 casos de cáncer contabilizados en dicho barrio constituyen un valor significativamente alto como para que la investigadora pueda concluir que existe algún factor especial en la comunidad bajo estudio que hace aumentar la probabilidad de que un niño padezca cáncer? O, por el contrario, ¿es posible que no haya nada especial en el barrio y que el mayor número de casos de cáncer se deba únicamente al azar? En este capítulo se verá cómo se pueden contestar las preguntas anteriores.

9.2 Contrastes de hipótesis y niveles de significación

La inferencia estadística es la ciencia que permite extraer conclusiones sobre una población a partir de la información contenida en una muestra. Un tipo especial de inferencia consiste en el contraste de hipótesis relativas a ciertos parámetros de la distribución poblacional. Por lo general, estas hipótesis establecerán que un parámetro poblacional, tal como la media o la varianza de la población, tiene un valor que cae dentro de una determinada región. En consecuencia, se deberá decidir si esta hipótesis es consistente con los datos observados en la muestra.

Definición

Una *hipótesis estadística* es una sentencia sobre la naturaleza de una población. Por lo general, se formula en términos de un determinado parámetro de la población.

Para contrastar una hipótesis estadística, se debe decidir si dicha hipótesis parece consistente con los datos de la muestra. Por ejemplo, supongamos que una compañía de tabaco mantiene que ha descubierto una nueva forma de secar las hojas de tabaco que garantiza que el contenido medio de nicotina por cigarrillo es menor o igual que 1,5 miligramos. Supongamos que un investigador es escéptico acerca de lo mantenido por la compañía, puesto que realmente cree que el contenido medio de nicotina excede de 1,5 miligramos. La hipótesis estadística que deberá contrastar, denominada *hipótesis nula* y que se denota como H_0, es que el contenido medio de nicotina por cigarrillo es menor o igual que 1,5 miligramos. Simbólicamente, si el contenido medio de nicotina se representa por μ, esta hipótesis nula se expresará como

$$H_0: \mu \leq 1{,}5$$

La alternativa a la hipótesis nula que el investigador está intentando contrastar se denomina *hipótesis alternativa* y se denota como H_1. En nuestro ejemplo, H_1 es la hipótesis de que el contenido medio de nicotina supera los 1,5 miligramos, que se denota por

$$H_1: \mu > 1{,}5$$

La *hipótesis nula*, denotada por H_0, es una sentencia acerca de un parámetro de la población. La hipótesis alternativa se denota por H_1. La hipótesis nula se rechazará si nos parece inconsistente con los datos muestrales, en caso contrario no se rechazará.

Para contrastar la hipótesis nula de que el contenido medio de nicotina por cigarrillo es menor o igual que 1,5 miligramos, se seleccionará una muestra aleatoria de cigarrillos producidos a partir de hojas secadas con el nuevo método y se medirán sus contenidos de nicotina. Si los datos muestrales resultantes no son "consistentes" con la hipótesis nula, se rechazará esta hipótesis; si, por el contrario, son "consistentes" con la hipótesis nula, ésta no será rechazada.

La decisión de rechazar o no rechazar la hipótesis nula se basa en el valor de un estadístico del contraste.

Definición

Un *estadístico del contraste* es un estadístico cuyo valor se determina a partir de los datos de la muestra. Dependiendo del valor que tome este estadístico del contraste, la hipótesis nula será o no rechazada.

En el ejemplo de los cigarrillos, el estadístico del contraste podría ser el contenido medio de nicotina de los cigarrillos de la muestra. El contraste estadístico rechazará la hipótesis nula cuando este estadístico del contraste sea suficientemente superior a 1,5. En general, si TS (de las siglas inglesas *Test Statistic*) denota el estadístico del contraste, para poder llevar a cabo el contraste se deberá especificar el conjunto de valores de TS para los que se deberá rechazar la hipótesis nula.

Definición

La *región crítica*, también denominada *región de rechazo*, es el conjunto de valores del estadístico del contraste para los que se rechazará la hipótesis nula.

El contraste estadístico de la hipótesis nula H_0 quedará completamente especificado en cuanto se determinen el estadístico del contraste y la región crítica. Si el estadístico del con-

traste se denota por TS y la región crítica por C, el contraste estadístico de la hipótesis nula actuará como sigue:

Rechazar H_0	si TS está en C
No rechazar H_0	si TS no está en C

En el ejemplo de los cigarrillos antes considerado, si se supiera que la desviación típica de los contenidos de nicotina de los cigarrillos es de 0,8 miligramos, una forma de contrastar la hipótesis nula podría consistir en utilizar $\overline{X}$, la media muestral de los contenidos de nicotina, como estadístico del contraste y como región crítica

$$C = \left\{\overline{X} \geq 1{,}5 + \frac{1{,}312}{\sqrt{n}}\right\}$$

Es decir, la hipótesis nula se deberá

Rechazar	si $\overline{X} \geq 1{,}5 + \dfrac{1{,}312}{\sqrt{n}}$
No rechazar	en otro caso,

siendo n el tamaño muestral. (El razonamiento por el que se ha elegido esta región crítica concreta se verá en el siguiente apartado del capítulo).

Por ejemplo, si se emplea la anterior forma de actuar frente al contraste, y el tamaño de la muestra fuera de 36, la hipótesis nula de que la media poblacional es menor o igual que 1,5 resultará rechazada si $\overline{X} \geq 1{,}719$ y no se rechazará si $\overline{X} < 1{,}719$. Es importante observar que, incluso aunque el valor estimado de μ –es decir, el valor de la media muestral $\overline{X}$– exceda de 1,5 puede que no se rechace la hipótesis nula. De hecho, cuando $n = 36$, un valor de la media muestral igual a 1,7 no producirá el rechazo de la hipótesis nula. Esto es cierto pese a que un valor tan alto de la media muestral parece no estar a favor de la hipótesis nula. Sin embargo, es consistente con la hipótesis nula en el sentido de que, si la media poblacional fuera de 1,5, existiría una razonable probabilidad de que la media de una muestra de tamaño 36 pudiera ser mayor o igual que 1,7. Por su parte, un valor de la media muestral mayor o igual que 1,9 es tan improbable que ocurra, si realmente fuera cierto que la media poblacional es menor o igual que 1,5, que nos debe llevar a rechazar la hipótesis nula.

El rechazo de la hipótesis nula H_0 es una sentencia fuerte en el sentido de que H_0 no parece ser consistente con los datos observados. No rechazar la hipótesis nula es una sentencia débil que se debería interpretar en el sentido de que H_0 es consistente con los datos

Así pues, en cualquier proceso de contraste de una determinada hipótesis nula se pueden cometer dos tipos de errores. El primero de ellos, llamado *error tipo I*, se produce

cuando se rechaza la hipótesis nula H_0 siendo ésta cierta. El segundo, denominado *error tipo II*, ocurre cuando no se rechaza la hipótesis nula H_0 siendo ésta falsa. Ahora bien, se ha de tener en mente que el objetivo de un contraste estadístico no consiste en determinar si H_0 es cierta, sino que por el contrario consiste en determinar si lo indicado por H_0 es consistente con los datos resultantes. Puesto que el objetivo es esto último, es razonable que H_0 se debería rechazar solamente cuando los datos muestrales son muy improbables de que ocurran si H_0 es cierta. La forma clásica de conseguir esto es la de especificar un pequeño valor α y, luego, obligar a que el contraste actúe de forma que, siempre que H_0 sea cierta, se cumpla que la probabilidad de rechazar H_0 sea menor o igual que α. El valor α, llamado *nivel de significación* del contraste, se suele fijar por adelantado, y sus valores habituales son $\alpha = 0{,}10$, 0,05 y 0,01.

El procedimiento clásico para contrastar una hipótesis nula consiste en fijar, primero, un pequeño nivel de significación α y, después, obligar a que, si H_0 es cierta, la probabilidad de rechazar H_0 sea menor o igual que α.

Debido a la asimetría de los contrastes con respecto a las hipótesis nula y alternativa, solamente se puede considerar que una hipótesis ha sido "probada" por los datos cuando la hipótesis nula se haya rechazado (y, por lo tanto, haya quedado "probado" que la hipótesis alternativa es cierta). Por esta razón, siempre se deberá tener en cuenta la siguiente regla.

Si uno está intentando probar una cierta hipótesis, está hipótesis deberá designarse como hipótesis alternativa. Del mismo modo, si uno intenta desacreditar una hipótesis, ésta deberá designarse como hipótesis nula.

Así, por ejemplo, si la compañía de tabaco está llevando a cabo el experimento con la intención de probar que el contenido medio en nicotina de sus cigarrillos es menor que 1,5, deberá tomar como hipótesis nula

$$H_0: \mu \geq 1{,}5$$

y como hipótesis alternativa

$$H_1: \mu < 1{,}5$$

De esta forma, la compañía podrá utilizar el rechazo de la hipótesis nula como "demostración" de que el contenido medio en nicotina es menor que 1,5 miligramos.

Supongamos ahora que se pretende llevar a cabo un contraste de cierta hipótesis referida a θ, un determinado parámetro de la población. En concreto, dada una región R, supon-

gamos que se está intentando contrastar la hipótesis nula de que θ esté en R. Esto es, se desea contrastar

$$H_0: \theta \text{ está contenido en } R$$

frente a la alternativa

$$H_1: \theta \text{ no está contenido en } R$$

Una forma para llevar a cabo el contraste de H_0, a nivel de significación α, empieza buscando un estimador puntual de θ, para después, a partir de él, actuar rechazando H_0 siempre que el valor de este estimador puntual se encuentre "muy alejado" de la región R. Para determinar lo alejado que ha de estar el estimador de la región R para que esté justificado rechazar H_0 se necesita, en primer lugar, determinar la distribución de probabilidad del estimador puntual cuando H_0 es cierta. Esto nos permitirá obtener una región crítica apropiada que cumpla que la probabilidad de que el estimador que caiga dentro de dicha región sea menor o igual que α, si H_0 es cierta. En la siguiente sección se explicará esta forma de actuar ante contrastes de hipótesis que afectan a la media de una población normal.

Problemas

1. Consideremos un juicio en el que el jurado debe decidir entre la hipótesis A, de que el acusado es culpable, y la hipótesis B, de que es inocente.

   (a) En el marco del contraste de hipótesis y teniendo en cuenta el sistema legal vigente, ¿qué hipótesis debería ser la hipótesis nula?

   (b) ¿Cuál debería ser el nivel de significación apropiado en esta situación?

2. La compañía farmacéutica británica Glaxo Holdings recientemente ha desarrollado un nuevo medicamento para la migraña (que produce un fuerte dolor de cabeza). Glaxo mantiene que este medicamento, llamado *somatriptan,* tarda menos de 10 minutos en ser absorbido por la sangre. Para convencer a la Administración Sanitaria de lo que mantiene, Glaxo ha llevado a cabo un experimento sobre un conjunto de pacientes elegido aleatoriamente. ¿Qué hipótesis debería tomar Glaxo como hipótesis nula y como hipótesis alternativa?

3. Supongamos que en un determinado contraste de

   $$H_0: \mu = 0 \quad \text{frente a} \quad H_1: \mu \neq 0$$

   resultó que H_0 fue rechazada, al nivel de significación del 5%. ¿Cuál o cuáles de las siguientes sentencias es o son correctas?:

   (a) Los datos prueban que μ es significativamente distinto de 0, lo que significa que se encuentra muy alejado de 0.

   (b) Los datos son significativamente fuertes para poder concluir que μ no es igual a 0.

   (c) La probabilidad de que μ sea igual a 0 es menor que 0,05.

   (d) Se ha rechazado la hipótesis de que μ es igual a 0 mediante un procedimiento por el que, si μ es igual a 0, solamente se rechaza H_0 en un 5% de los casos.

(Public Information Office, University of California, Berkeley)

Jerzy Neyman

Perspectiva histórica

El concepto de nivel de significación se debió originariamente al estadístico inglés Ronald A. Fisher. Éste igualmente formuló el concepto de hipótesis nula como aquella que uno intenta desacreditar. En palabras de Fisher: "Puede decirse que todos los experimentos se diseñan para poder asignar una probabilidad al hecho de que los resultados se opongan a la hipótesis nula." La idea de la hipótesis alternativa se debe a los trabajos conjuntos del estadístico de origen polaco Jerzy Neyman y de su habitual colaborador por muchos años, Egon Pearson (hijo de Karl). Fisher, sin embargo, no aceptó la idea de especificar una hipótesis alternativa, arguyendo que en la mayoría de las aplicaciones científicas no era posible especificar las alternativas; de esta forma se produjo una gran disputa entre Fisher, por un lado, y Neyman y Pearson, por otro. Debido tanto al temperamento de Fisher, al que no le gustaba demasiado explicar las cosas, como al hecho de que éste ya mantenía una discusión previa con Neyman sobre los beneficios relativos de los estimadores por intervalos de confianza, propuestos por Neyman, frente a los estimadores de confianza fiduciarios de Fisher (hoy en desuso), la disputa se convirtió en extremadamente personal y mordaz. En una ocasión, Fisher calificó la posición de Neyman como "terrorífica para la libertad intelectual en el mundo occidental".

Fisher es famoso por sus disputas científicas. Aparte de las que se acaban de comentar, mantuvo también un aún más acalorado debate con Karl Pearson acerca de los méritos de dos procedimientos diferentes para obtener estimadores puntuales, conocidos como el *método de los momentos* y el *método de máxima verosimilitud*, respectivamente. Fisher, que fue el fundador del área de la genética de poblaciones, también discutió durante mucho tiempo con Sewell Wright, otro influyente genetista de poblaciones, acerca del papel desempeñado por el azar en la determinación de las frecuencias de genes futuras. (Es curioso que fuera el biólogo Wright quién mantuvo que la causalidad era el factor clave en los procesos de evolución a largo plazo.)

4. Denotemos por μ el valor medio de una determinada población. Supongamos que, para contrastar

$$H_0: \mu \leq 1{,}5$$

frente a

$$H_1: \mu > 1{,}5$$

se ha seleccionado una muestra de la citada población.

(a) Si no se ha podido rechazar H_0 con dicha muestra, ¿significa esto que, en caso de haberse planteado el contraste de

$$H_0: \mu > 1{,}5 \quad \text{frente a} \quad H_1: \mu \leq 1{,}5$$

sí que se hubiera rechazado la hipótesis nula con la muestra extraída.

(b) Supongamos que se hubiera rechazado la hipótesis H_0 del contraste inicial. ¿Implica esto que, ante el contraste de

$$H_0: \mu > 1{,}5 \quad \text{frente a} \quad H_1: \mu \leq 1{,}5$$

no se habría rechazado la hipótesis nula, con la misma muestra?

Explique las respuestas, si se asume que todos los contrastes se realizan al nivel de significación del 5%.

9.3 Contrastes relativos a la media de una población normal: caso de varianza conocida

Supongamos que $X_1, \ldots, X_n$ es una muestra aleatoria procedente de una población normal con media desconocida μ y varianza conocida σ^2; adicionalmente, supongamos que se pretende contrastar la hipótesis nula de que la media μ es igual a un determinado valor frente a la hipótesis alternativa de que μ no es igual a dicho valor. Es decir, se desea contrastar

$$H_0: \mu = \mu_0$$

frente a la hipótesis alternativa

$$H_1: \mu \neq \mu_0$$

siendo μ_0 un valor prefijado de antemano.

Dado que el estimador puntual natural de la media poblacional μ es la media muestral

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

parece razonable rechazar la hipótesis nula de que la media poblacional es igual a μ_0 cuando $\overline{X}$ se separa mucho de μ_0. Esto es, la región crítica del contraste debería adoptar la forma

$$C = \{X_1, \cdots, X_n: \quad |\overline{X} - \mu_0| \geq c\}$$

para algún valor apropiado de c.

Supongamos que se quiere que el contraste tenga un nivel de significación α. En este caso, el valor de c se elegirá de forma que, cuando la media poblacional sea igual a μ_0, la probabilidad de que $\overline{X}$ difiera de μ_0 en un valor mayor o igual que c sea igual a α. Es decir, c debe verificar que

$$P\{|\overline{X} - \mu_0| \geq c\} = \alpha \quad \text{cuando } \mu = \mu_0 \tag{9.1}$$

Ahora bien, cuando μ es igual a μ_0, $\overline{X}$ se distribuye según una normal de media μ_0 y desviación típica $\sigma/\sqrt{n}$; en consecuencia, la variable estandarizada Z, definida por

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{\sqrt{n}}{\sigma}(\overline{X} - \mu_0)$$

seguirá una distribución normal estándar. Dado que la desigualdad

$$|\bar{X} - \mu_0| \geq c$$

es equivalente a

$$\frac{\sqrt{n}}{\sigma}|\bar{X} - \mu_0| \geq \frac{\sqrt{n}}{\sigma}c$$

se tiene que la probabilidad que figura en (9.1) y equivale a

$$P\{|Z| \geq \sqrt{n}\,\frac{c}{\sigma}\} = \alpha$$

Si ahora se tiene en cuenta que la probabilidad de que el valor absoluto de una normal estándar sobrepase un determinado valor es igual al doble de la probabilidad de que una normal estándar sea mayor que el valor citado (véase la figura 9.1), se desprende de lo anterior que

$$P\left\{Z \geq \sqrt{n}\frac{c}{\sigma}\right\} = \alpha$$

o

$$P\left\{Z \geq \sqrt{n}\frac{c}{\sigma}\right\} = \frac{\alpha}{2}$$

Finalmente, puesto que $z_{\alpha/2}$ se define de forma que se cumpla que

$$P\{Z \geq z_{\alpha/2}\} = \frac{\alpha}{2}$$

se desprende de todo lo anterior que

$$\sqrt{n}\frac{c}{\sigma} = z_{\alpha/2}$$

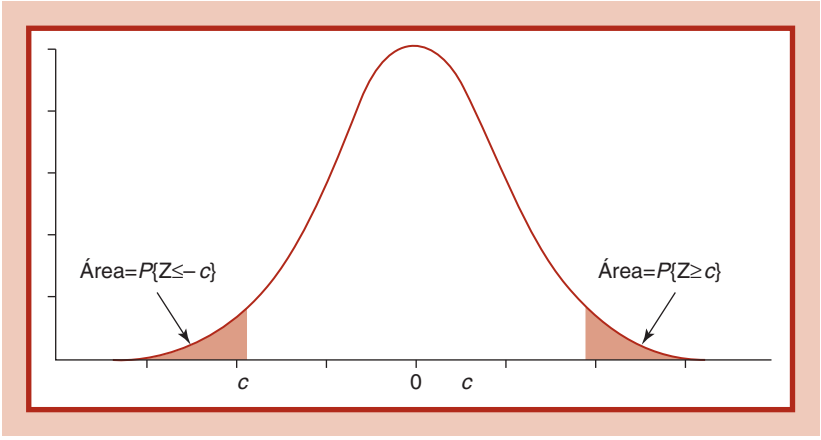


Figura 9.1 $P\{|Z| \geq c\} = P\{Z \geq c\} + P\{Z \leq -c\} = 2P\{Z \geq c\}$.

o

$$c = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Por consiguiente, al nivel de significación α, el contraste de la hipótesis nula de que la media poblacional es igual al valor prefijado μ_0 frente a la hipótesis alternativa de que la media anterior no es igual a μ_0 rechazará la hipótesis nula si

$$|\bar{X} - \mu_0| \geq z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

o, equivalentemente, se deberá

Rechazar H_0 si	si $\dfrac{\sqrt{n}}{\sigma}|\bar{X} - \mu_0| \geq z_{\alpha/2}$
No rechazar H_0	en otro caso

La forma de actuar de este contraste se representa gráficamente en la figura 9.2. Observe que en esta figura también se ha representado sobre la línea horizontal la función de densidad de la normal estándar, puesto que ésta es la densidad del estadístico del contraste $\sqrt{n}(\bar{X} - \mu_0)/\sigma$ cuando H_0 es cierta. Adicionalmente, debido a este hecho, el contraste anterior se denomina *contraste de la Z*.

Ejemplo 9.1 Supongamos que, si una determinada estrella emite una señal de intensidad μ, el valor recibido en un observatorio terrestre es una variable aleatoria normal con media μ y desviación típica 4. En otras palabras, el valor de la señal emitida se ve alterado por un *ruido aleatorio*, que se distribuye según una normal con media 0 y desviación típica 4. Se sospecha que la intensidad de la señal es igual a 10. Contraste si esta hipótesis sería plausible si la misma señal se recibiera independientemente 20 veces y la media de los 20 valores recibidos fuera igual a 11,6. Utilice un nivel de significación del 5%.

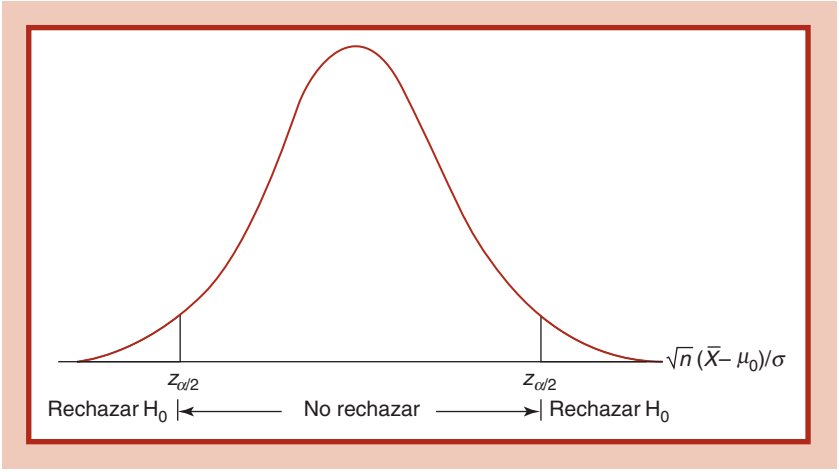


Figura 9.2 Contraste de H_0: $\mu = \mu_0$ frente a H_1: $\mu \neq \mu_0$.

Solución Si μ representa la intensidad real de la señal emitida, la hipótesis nula que se desea contrastar es

$$H_0: \mu = 10$$

frente a la alternativa

$$H_1: \mu \neq 10$$

Supongamos que se intenta levar a cabo el contraste al nivel de significación del 5%. Se debe empezar calculando el valor del estadístico

$$\frac{\sqrt{n}}{\sigma}|\bar{X} - \mu_0| = \frac{\sqrt{20}}{4}|11{,}6 - 10| = 1{,}79$$

Dado que este valor es menor que $z_{0,025} = 1{,}96$, la hipótesis nula no resulta rechazada. En otras palabras, se concluye que los datos no son inconsistentes con la hipótesis nula de que el valor de la señal es igual a 10. La justificación de este aserto se basa en el hecho de que se puede observar una media muestral que diste de 10 al menos tanto como la media observada, cuando H_0 es cierta, en más de un 5% de los casos. Observe, sin embargo, que si el nivel de significación elegido hubiera sido $\alpha = 0{,}1$, a diferencia del anterior $\alpha = 0{,}05$, la hipótesis nula se hubiera rechazado (puesto que $z_{\alpha/2} = z_{0,05} = 1{,}645$). ■

Es importante observar que el nivel de significación "correcto" que se debe utilizar en un determinado contraste de hipótesis depende de las circunstancias particulares bajo las que se planteó dicho contraste. Si el hecho de rechazar la hipótesis nula implicara un gran coste, que se ahorraría si H_0 fuera cierta, posiblemente decidiríamos actuar conservadoramente y elegiríamos un pequeño nivel de significación. Por ejemplo, supongamos que H_1 coincide con la hipótesis de que se prefiere un nuevo método de producción frente al que se utiliza actualmente. Dado que rechazar H_0 implicaría aceptar un cambio en el proceso de producción es natural que se desee asegurar que, cuando H_0 sea cierta, la probabilidad de rechazar H_0 sea pequeña; esto es, deberíamos elegir un valor de α pequeño. De igual forma, si se creyera claramente que la hipótesis nula es cierta, deberíamos demandar una alta evidencia de los datos en sentido contrario para rechazar H_0; en consecuencia, se debería elegir un nivel de significación muy pequeño.

El contraste de hipótesis que se acaba de analizar se puede describir como sigue: Primero se calcula el valor del estadístico del contraste $\sqrt{n}(\bar{X} - \mu_0)/\sigma$. Si este valor es v, se rechazará H_0 si la probabilidad de que el estadístico del contraste en valor absoluto sea mayor o igual que $|v|$ es menor o igual que α, cuando H_0 es cierta. Por consiguiente, se deberán computar primero el valor v del estadístico del contraste y después la probabilidad de que una normal estándar en valor absoluto supere $|v|$. Esta probabilidad, llamada *p valor*, proporciona un nivel de significación crítico, en el sentido de que H_0 se rechazará si el *p* valor es menor o igual que el nivel de significación α prefijado para el contraste, y no se rechazará en el caso contrario.

El *p* valor es el menor nivel de significación al que se debería rechazar la hipótesis nula con los datos disponibles. Se puede interpretar como la probabilidad de que se puedan obtener

unos datos que se manifiesten en contra de H_0 al menos tanto como los datos observados. Un p valor pequeño (de 0,05 o menos) es un fuerte indicador de que la hipótesis nula no es cierta. Cuanto menor es el p valor, mayor es la evidencia sobre la falsedad de H_0.

En la práctica ocurre muy a menudo que el nivel de significación no se fija de antemano, sino que por el contrario se utilizan los datos para obtener el p valor. Por lo general, este valor es o muy grande, en cuyo caso está claro que la hipótesis nula no se debe rechazar, o muy pequeño, lo que indica claramente que se deberá rechazar la hipótesis nula.

Ejemplo 9.2 Supongamos que la media de los 20 valores del ejemplo 9.1 fuera igual a 10,8. En este caso, el valor absoluto del estadístico del contraste sería

$$\frac{\sqrt{n}}{\sigma}|\bar{X} - \mu_0| = \frac{\sqrt{20}}{4}|10{,}8 - 10| = 0{,}894$$

Dado que

$$\begin{aligned} P\{|Z| \geq 0{,}894\} &= 2P\{Z \geq 0{,}894\} \\ &= 0{,}371 \text{ (obtenido a partir de la tabla D.1),} \end{aligned}$$

el p valor resultante es igual a 0,371. En consecuencia, la hipótesis nula de que el valor de la señal es 10 no deberá rechazarse a ningún nivel de significación inferior a 0,371. Puesto que no se usan niveles de significación tan altos como este valor, no se rechazará la hipótesis nula.

Por otro lado, si el valor de la media muestral fuera 7,8, se obtendría como valor absoluto del estadístico del contraste

$$\frac{\sqrt{20}}{4}(2{,}2) = 2{,}46$$

y, por consiguiente, el p valor resultante sería

$$\begin{aligned} p \text{ valor} &= P\{|Z| \geq 2{,}46\} \\ &= 2P\{Z \geq 2{,}46\} \\ &= 0{,}014 \end{aligned}$$

Así pues, se debería rechazar H_0 con cualquier nivel de significación superior a 0,014, mientras que con niveles de significación inferiores a este valor no se debería rechazar la hipótesis nula. ■

En el siguiente ejemplo se trata de determinar la probabilidad de no rechazar la hipótesis nula cuando ésta es falsa.

Ejemplo 9.3 En el ejemplo 9.1, al nivel de significación de 0,05, ¿cuál es la probabilidad de que no se rechace la hipótesis nula (de que la intensidad de la señal sea igual a 10) si el valor real de la señal fuera de 9,2?

Solución En el ejemplo 9.1, $\sigma = 4$ y $n = 20$. Por consiguiente, para contrastar

$$H_0: \mu = 10 \quad \text{frente a} \quad H_1: \mu \neq 10$$

se deberá rechazar H_0 si

$$\frac{\sqrt{20}}{4}|\overline{X} - 10| \geq z_{0,025}$$

o, equivalentemente, si

$$|\overline{X} - 10| \geq \frac{4z_{0,025}}{\sqrt{20}}$$

Puesto que $4z_{0,025}/\sqrt{20} = 4 \times 1{,}96/\sqrt{20} = 1{,}753$, esto significa que H_0 se deberá rechazar si la distancia entre $\overline{X}$ y 10 es como mínimo de 1,753. Es decir, se rechazará H_0 tanto si

$$\overline{X} \geq 10 + 1{,}753$$

como si

$$\overline{X} \leq 10 - 1{,}753$$

Es decir, si

$$\overline{X} \geq 11{,}753 \quad \text{o} \quad \overline{X} \leq 8{,}247$$

H_0 será rechazada.

Ahora bien, si la media poblacional es de 9,2, $\overline{X}$ se distribuye según una normal con media de 9,2 y una desviación típica de $4/\sqrt{20} = 0{,}894$; así pues, la variable estandarizada

$$Z = \frac{\overline{X} - 9{,}2}{0{,}894}$$

será una normal estándar. Por consiguiente, si el valor de la señal es de 9,2, se ve que

$$\begin{aligned} P\{\text{rechazar } H_0\} &= P\{\overline{X} \geq 11{,}753\} + P\{\overline{X} \leq 8{,}247\} \\ &= P\left\{\frac{\overline{X} - 9{,}2}{0{,}894} \geq \frac{11{,}753 - 9{,}2}{0{,}894}\right\} + P\left\{\frac{\overline{X} - 9{,}2}{0{,}894} \leq \frac{8{,}247 - 9{,}2}{0{,}894}\right\} \\ &= P\{Z \geq 2{,}856\} + P\{Z \leq -1{,}066\} \\ &= 0{,}0021 + 0{,}1432 \\ &= 0{,}1453 \end{aligned}$$

Es decir, si el valor de la señal fuera de 9,2, existe una probabilidad del 85,47% de que el contraste, al nivel de significación de 0,05, no rechace la hipótesis nula de que el valor de la señal es igual a 10. ■

Problemas

Asuma en todos los problemas que la distribución subyacente a los datos es normal.

1. El aparato que utiliza un astrónomo para medir distancias realiza mediciones que tienen una media igual a la distancia real al cuerpo observado con una desviación típica de 0,5 años luz. Se piensa que la distancia de la tierra al asteroide phyla es de 14,4 años luz. Al nivel de significación del 5%, contraste esta hipótesis si, tras realizar 6 mediciones independientes, el astrónomo obtuvo los datos siguientes:

   15,1, 14,8, 14,0, 15,2, 14,7, 14,5

2. Con una muestra de peces del Lago Michigan extraída hace tiempo se obtuvo que la media de las concentraciones de bifenil policlorinado (BPC) por pez fue de 11,2 partes por millón, con una desviación típica de 2 partes por millón. Recientemente, se ha extraído una nueva muestra de 10 peces cuyas concentraciones de BPC fueron:

   11,5, 12,0, 11,6, 11,8, 10,4, 10,8, 12,2, 11,9, 12,4, 12,6

   Suponiendo que la desviación típica continúa siendo igual a 2 partes por millón, contraste la hipótesis de que la actual concentración media de BPC no ha cambiado y de que sigue con un nivel de 11,2 partes por millón. Utilice un nivel de significación del 5%.

3. Se ha seleccionado una muestra de tamaño 9 para contrastar las hipótesis

   $$H_0: \mu = 105 \qquad \text{frente a} \qquad H_1: \mu \neq 105$$

   Si la media muestral resultante fue $\overline{X} = 100$, calcule el p valor si se sabe que la desviación típica poblacional es

   (a) $\sigma = 5$

   (b) $\sigma = 10$

   (c) $\sigma = 15$

   ¿En qué casos se deberá rechazar la hipótesis nula, al nivel de significación del 5%? ¿Y con un nivel del 1%?

4. Repita el problema 3 si se asume que la media muestral es la misma pero que el tamaño muestral es 36.

5. Una colonia de ratones de laboratorio se compone de varios miles de unidades. El peso medio del total de los ratones es de 32 gramos con una desviación típica de 4 gramos. Un científico pide a un empleado del laboratorio que seleccione 25 ratones para llevar a cabo un determinado experimento. Sin embargo, antes de realizarlo, el científico decide pesar los ratones para comprobar que la selección hecha por el empleado realmente fue aleatoria y no se cometió ningún tipo de sesgo inconsciente (por ejemplo, que el empleado no hubiera seleccionado los ratones más lentos, lo cual podría indicar algún tipo de inferioridad dentro del grupo). Si la media muestral de los pesos de los 25 ratones selec-

cionados fue 30,4, al nivel de significación del 5%, ¿este dato manifiesta una evidencia suficiente en contra de la hipótesis de que la selección se llevó a cabo de forma aleatoria?

6. Se sabe que el valor de recepción de una señal en una estación local es igual al valor emitido más un error aleatorio con media 0 y desviación típica 2. Si el mismo valor se emite 7 veces, calcule el p valor al contrastar la hipótesis nula de que el valor emitido fue igual a 14, si los valores recibidos resultaron ser:

   14,6, 14,8, 15,1, 13,2, 12,4, 16,8, 16,3

7. Los datos históricos indican que la cantidad de agua usada diariamente en los hogares de Estados Unidos tiende a distribuirse según una normal con media 360 galones y desviación típica 40 galones. Para ver si estos valores todavía se mantienen en la actualidad se seleccionó una muestra aleatoria de 200 hogares. El consumo medio diario de agua en esos hogares fue de 374 galones.

   (a) ¿Estos datos están en concordancia con la distribución histórica? Utilice un nivel de significación del 5%.

   (b) ¿Cuál es el p valor resultante?

8. Cuando un proceso de producción funciona adecuadamente, los elementos producidos presentan una característica mesurable cuya media es 122 y con una desviación típica de 9. Sin embargo, si el proceso se desajusta y se sale de control, se produce un cambio en la media de la dicha característica de las unidades producidas. Contraste la hipótesis de que el proceso está actualmente bajo control si se ha seleccionado una muestra aleatoria de 10 unidades producidas recientemente y sus medidas de la característica han sido:

   123, 120, 115, 125, 131, 127, 130, 118, 125, 128

   Indique la hipótesis nula y la alternativa, y calcule el p valor.

9. Una compañía de *leasing* opera bajo la hipótesis de que la cantidad de millas recorridas por los coches bajo *leasing* se distribuye según una normal de media 13 500 y desviación típica 4000 millas. Para comprobar si esta hipótesis es válida se seleccionó una muestra aleatoria de 36 coches con un año de antigüedad. Si la media de las millas recorridas por esos 36 coches fue de 15 233, ¿qué conclusión se puede extraer?

10. Se sabe que la distribución de una determinada población tiene una desviación típica de 20. Calcule el p valor al contrastar la hipótesis de que la media poblacional es igual a 50, si la media de una muestra de 64 observaciones resultó ser:

    (a) 52,5

    (b) 55,0

    (c) 57,5

11. Las autoridades de tráfico mantienen que los semáforos están en rojo durante un tiempo que se distribuye según una normal de media 30 segundos y desviación típica

1,4 segundos. Para contrastar esta afirmación se comprobó una muestra de 40 semáforos. Si el tiempo medio que estos semáforos estuvieron en rojo es igual a 32,2 segundos, ¿se puede concluir, al nivel de significación del 5%, que las autoridades están equivocadas? ¿Y si el nivel de significación es del 1%?

12. El número de casos de cáncer infantil que se presentan durante un periodo de 3 años en una comunidad de un determinado tamaño sigue aproximadamente una distribución normal con media 16,2 y desviación típica 4,7. Para analizar si esta distribución cambia cuando la comunidad está situada en las proximidades de un campo electromagnético elevado, un investigador seleccionó una comunidad próxima a un campo de este tipo y averiguó después que se habían producido un total de 32 casos de cáncer infantil en los últimos 3 años. Con estos datos calcule el p valor al contrastar la hipótesis de que la distribución del número de casos de cáncer infantil que ocurren en las comunidades próximas a altos campos electromagnéticos se mantiene normal con media 16,2 y desviación típica 4,7.

13. Se sabe que los datos siguientes provienen de una población normal con desviación típica 2. Utilícelos para contrastar la hipótesis nula de que la media poblacional es igual a 15. Determine los niveles de significación a los que se debe rechazar y no rechazar la hipótesis anterior.

    15,6, 16,4, 14,8, 17,2, 16,9, 15,3, 14,0, 15,9

14. Supongamos que, en el problema 1, la distancia real al asteroide phyla es de 14,8 años luz. Con una serie de 10 mediciones, cada una de las cuales se distribuye según una normal de media igual a la distancia real y desviación típica 0, 8 años luz, ¿cuál es la probabilidad de que se llegue a rechazar la hipótesis nula de que la distancia al asteroide coincide con 14 años luz? Utilice un nivel de significación del 1%.

15. En el problema 6 calcule la probabilidad de que, al nivel de significación del 5%, se rechace la hipótesis nula de que el valor emitido fue 14, si el valor realmente emitido fue:

    (a) 15

    (b) 13

    (c) 16

9.3.1 Contrastes unilaterales

Hasta este momento se han considerado contrastes con hipótesis alternativa bilateral, en los que la hipótesis nula consistía en que μ coincide con μ_0. En esa situación se debía rechazar la hipótesis nula siempre que $\overline{X}$ fuera mucho mayor o mucho menor que μ_0. Sin embargo, en muchos casos, la hipótesis que se desea contrastar es que la media sea menor o igual que un determinado valor μ_0 frente a la hipótesis alternativa de que la media sea mayor que dicho valor. Es decir, en ocasiones uno está interesado en contrastar

$$H_0: \mu \leq \mu_0$$

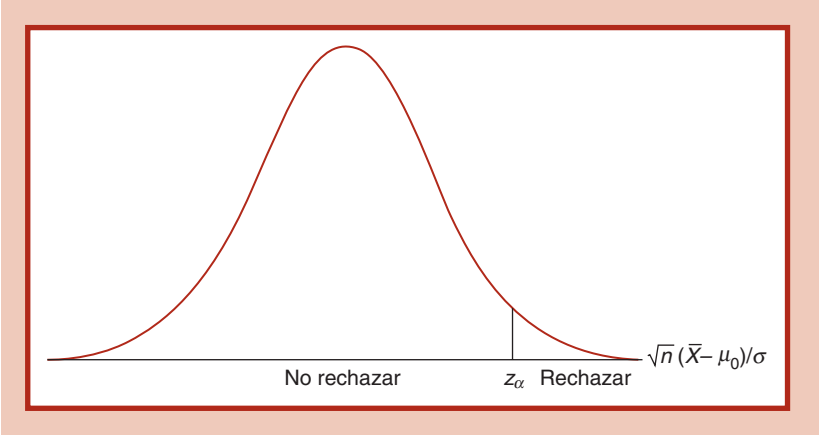


Figura 9.3 Contrastando H_0: $\mu \leq \mu_0$ frente a H_1: $\mu > \mu_0$.

frente a la alternativa

$$H_1: \mu > \mu_0$$

Puesto que lógicamente uno desearía rechazar la hipótesis nula cuando la media muestral $\bar{X}$ fuera mucho mayor que μ_0 (y no cuando fuera mucho menor), se puede demostrar exactamente de la misma forma que se hizo en el caso bilateral que el contraste, al nivel de significación α, actuará como sigue:

$$\begin{array}{ll} \text{Rechazar } H_0 & \text{si } \sqrt{n}\,\dfrac{\bar{X}-\mu_0}{\sigma} \geq z_\alpha \\ \text{No rechazar } H_0 & \text{en caso contrario} \end{array}$$

Gráficamente esto se muestra en la figura 9.3.

Este contraste se puede llevar a cabo calculando primero el valor del estadístico del contraste $\sqrt{n}(\bar{X}-\mu_0)/\sigma$ y, después, obteniendo el p valor, que es igual a la probabilidad de que una variable aleatoria normal estándar sea mayor o igual que el valor obtenido del estadístico del contraste. Es decir, si este último valor es v,

$$p \text{ valor} = P\{Z \geq v\}$$

La hipótesis nula se deberá rechazar a cualquier nivel de significación mayor o igual que el p valor.

De forma similar, se puede contrastar la hipótesis nula de que

$$H_0: \mu \geq \mu_0$$

frente a la alternativa

$$H_1: \mu < \mu_0$$

calculando primero el valor del estadístico del contraste $\sqrt{n}(\bar{X}-\mu_0)/\sigma$ y, después, obteniendo el p valor que coincide con la probabilidad de que una variable aleatoria normal

estándar sea menor o igual que el valor obtenido del estadístico del contraste. La hipótesis nula se deberá rechazar si el nivel de significación es mayor o igual que el p valor.

Ejemplo 9.4 Para el conjunto de todos los cigarrillos comercializados actualmente, el contenido medio de nicotina por cigarrillo es mayor o igual que 1,5 miligramos. Una compañía de tabaco mantiene que ha descubierto una nueva técnica de curación de las hojas de tabaco que hace que el contenido medio de nicotina por cigarrillo sea menor que 1,5 miligramos. Para contrastar esta afirmación se analizó una muestra de 20 cigarrillos producidos con hojas curadas con la nueva técnica. Si se sabe que la desviación típica de los contenidos de nicotina es de 0,7 miligramos, ¿qué conclusiones se podrían sacar, al nivel de significación del 5%, si el contenido medio de nicotina para estos 20 cigarrillos resultó ser de 1,42 miligramos?

Solución Para contrastar si estos datos corroboran lo que mantiene la compañía se debería ver si se rechaza la hipótesis nula de que los cigarrillos curados con la nueva técnica no tienen un contenido medio en nicotina inferior a 1,5 miligramos. Es decir, se debería contrastar

$$H_0: \mu \geq 1{,}5$$

frente a lo que mantiene la compañía

$$H_1: \mu < 1{,}5$$

Puesto que el valor del estadístico del contraste es

$$\sqrt{n}\,\frac{\bar{X} - \mu_0}{\sigma} = \sqrt{20}\,\frac{1{,}42 - 1{,}5}{0{,}7} = -0{,}511$$

se tiene que el p valor resultante es

$$p \text{ valor} = P\{Z \leq -0{,}511\} = 0{,}305$$

Dado que este p valor es superior a 0,05, los anteriores datos no nos permiten rechazar la hipótesis nula y concluir que el contenido medio en nicotina por cigarrillo es menor que 1,5 miligramos. En otras palabras, pese a que la evidencia se muestra a favor de lo que mantiene la compañía (puesto que el contenido medio de nicotina de los cigarrillos de la muestra es realmente inferior a 1,5 miligramos), esta evidencia no es lo suficientemente fuerte como para *probar* lo indicado. Esto se debe a que un resultado que esté al menos tan a favor de de la hipótesis alternativa H_1 como la observada, se puede esperar que ocurra en un 30,5% de los casos, si el contenido medio de nicotina fuera de 1,5 miligramos por cigarrillo. ■

Un contraste de hipótesis estadístico en el que tanto la hipótesis nula como la hipótesis alternativa establecen que un parámetro es mayor (o menor) que un determinado valor se denominan contrastes *unilaterales*.

Hasta ahora se ha venido asumiendo que la distribución poblacional subyacente es normal. Sin embargo, tan solo se ha utilizado esta hipótesis para concluir que $\sqrt{n}(\bar{X} - \mu)/\sigma$

Tabla 9.1 Contrastes de hipótesis relativos a la media μ de una población normal con varianza conocida σ^2.

$X_1, \ldots, X_n$ *son los datos muestrales, y*

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

H_0	H_1	Estadístico del contraste TS	Contraste a nivel de significación α	p valor si TS $= \nu$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\sqrt{n}\dfrac{\overline{X} - \mu_0}{\sigma}$	Rechazar H_0 si $\lvert TS \rvert \geq z_{\alpha/2}$ No rechazar H_0 en otro caso	$2P\{Z \geq \lvert\nu\rvert\}$
$\mu \leq \mu_0$	$\mu > \mu_0$	$\sqrt{n}\dfrac{\overline{X} - \mu_0}{\sigma}$	Rechazar H_0 si TS $\geq z_\alpha$ No rechazar H_0 en otro caso	$P\{Z \geq \nu\}$
$\mu \geq \mu_0$	$\mu < \mu_0$	$\sqrt{n}\dfrac{\overline{X} - \mu_0}{\sigma}$	Rechazar H_0 si TS $\leq -z_\alpha$ No rechazar H_0 en otro caso	$P\{Z \leq \nu\}$

sigue una distribución normal estándar. Ahora bien, el teorema central del límite garantiza que este mismo resultado se verifica en términos aproximados cualquiera que sea la distribución poblacional subyacente, siempre que n sea razonablemente grande. Con carácter general, un tamaño muestral $n \geq 30$ suele ser suficiente. De hecho, para una gran cantidad de distribuciones poblacionales se pueden conseguir buenas aproximaciones a la normal con unos valores de n tan pequeños como 4 o 5. En consecuencia se pueden utilizar todos los contrastes de hipótesis desarrollados anteriormente incluso aunque la distribución poblacional subyacente no sea normal.

En la tabla 9.1 se resumen los contrastes analizados en este apartado.

Problemas

1. Los pesos de los salmones criados en una piscifactoría se distribuyen normalmente con una desviación típica de 1,2 libras. La piscifactoría mantiene que el peso medio de los salmones criados este año es como mínimo de 7,6 libras. Supongamos que el peso medio de una muestra aleatoria de 16 salmones resultó ser de 7,2 libras. Este valor ofrece una evidencia suficiente para que se rechace lo mantenido por la piscifactoría:

   (a) Al nivel de significación del 5%.

   (b) Al nivel de significación del 1%.

   (c) ¿Cuál es el p valor resultante?

2. Consideremos el contraste de H_0: $\mu \leq 100$ frente a H_1: $\mu > 100$ y supongamos que se ha extraído una muestra aleatoria de tamaño 20 cuya media muestral resultó ser $\overline{X} = 105$. Calcule el p valor del contraste si se sabe que la desviación típica poblacional es igual a:

   (a) 5

   (b) 10

   (c) 15

La Estadística en perspectiva

La Isla Three Mile

Aún hoy se cuestiona si el accidente nuclear de la Isla Three Mile, en el que se produjo un escape de radiación nuclear de bajo nivel, es la causa de un incremento en el número de casos de hipertiroidismo. Si no se trata rápidamente, el hipertiroidismo, que resulta cuando la glándula tiroides no funciona correctamente, puede ocasionar un retraso mental. Se sabe por los registros que, entre el 28 de marzo de 1979 (fecha del accidente) y el 28 de diciembre de 1979 (nueve meses más tarde), nacieron 11 bebés que sufrieron de hipertiroidismo en los alrededores de la central nuclear. Por los registros se sabe también que el número habitual de bebés con hipertiroidismo nacidos en la zona a lo largo de un periodo de 9 meses sigue aproximadamente una distribución normal con media 3 y desviación típica 2. Con esta información, empecemos calculando la probabilidad de que los 11 casos de hipertiroidismo registrados tras el accidente hayan ocurrido por azar.

Observe en primer lugar que, si el accidente no hubiera tenido efectos sobre la salud y, en consecuencia, los 9 meses posteriores al accidente hubieran sido meses normales, el número de bebés que padecieron hipertiroidismo debería seguir la citada distribución normal con media 3 y desviación típica 2. Por otra parte, si el accidente hubiera tenido un efecto nocivo sobre el hipertiroidismo, la media de la distribución debeía ser mayor que 3. Así pues, supongamos que los datos proceden de una distribución con una desviación típica 2 y llevemos a cabo el contraste de

$$H_0: \mu \leq 3 \qquad \text{frente a} \qquad H_1: \mu > 3$$

siendo μ el número medio de bebés con hipertiroidismo.

Dado que se registraron 11 casos, el p valor resultante es

$$\begin{aligned} p \text{ valor} &= P\{X \geq 11\} \\ &= P\{X \geq 10{,}5\}_{\text{corrección por continuidad}} \\ &= P\left\{\frac{X-3}{2} \geq \frac{10{,}5-3}{2}\right\} \\ &\approx P\{Z \geq 3{,}75\} \\ &< 0{,}0001 \end{aligned}$$

Así pues, se debe rechazar la hipótesis nula al nivel de significación del 1% (e incluso del 0,1%).

La Estadística en perspectiva (continuación)

Es importante destacar que este contraste *no demuestra* que el accidente nuclear fue la causa del incremento de los casos de hipertiroidismo; de hecho, *ni siquiera demuestra que hubo un aumento de esta enfermedad*. Realmente es difícil saber qué se puede concluir a partir de este contraste sin disponer de mucha más información. Por ejemplo, una dificultad surge de nuestra falta de información acerca de las razones que indujeron a analizar las hipótesis particulares citadas. Esto es, ¿existió alguna causa científica anterior para creer que la liberación de radiación nuclear podía ocasionar un aumento de hipertiroidismo en los recién nacidos, o simplemente alguien pensó en todas las posibles enfermedades que podían verse afectadas (posiblemente para una gran variedad de grupos de edad) y posteriormente contrastó si se produjeron cambios significativos en sus incidencias tras el accidente? El problema con tal enfoque (a menudo conocido como *minería de datos*, o *yendo de pesca*) es que, incluso aunque el accidente no hubiera producido cambios, podría ocurrir que, simplemente por azar, algunos de los muchos contrastes realizados resultaran significativos a favor del cambio. [Por ejemplo, si se llevaran a cabo 20 contrastes independientes de hipótesis, incluso aunque todas las hipótesis nulas fueran ciertas, al menos una de ellas sería rechazada, al nivel de significación del 1%, con una probabilidad $1 - (0{,}99)^{20} = 0{,}18$.]

Otra dificultad en la interpretación de nuestro contraste de hipótesis afecta a la confianza que se tiene en las cifras dadas. Por ejemplo, ¿podemos estar seguros de que, bajo condiciones de normalidad, el número medio de recién nacidos con hipertiroidismo es igual a 3? ¿No sería más verosímil decir que, aunque en media se diagnosticaran 3 casos de hipertiroidismo en los recién nacidos, podrían no haberse detectado otros casos existentes? ¿No podría ocurrir que la posibilidad de que existan casos no detectados haya sido menor tras el accidente, dado que todo el mundo estaba alerta de los posibles incrementos? Quizá también existieran distintos grados de hipertiroidismo, y un recién nacido al que se le diagnosticó la enfermedad en los tensos meses tras el accidente no habría tenido el mismo diagnóstico en tiempos normales.

Se ha de tener en cuenta que no estamos intentando razonar que no existió un incremento real en los casos de hipertiroidismo tras el accidente de la isla Three Mile. Por el contrario, estamos intentando que el lector sea consciente de las dificultades potenciales que existen cuando los resultados de un estudio estadístico se intentan evaluar correctamente.

3. Repita el problema 2, si se supone en esta ocasión que el valor de la media muestral es 108.

4. En ciertos procesos químicos es muy importante que una solución usada como reactivo tenga un nivel de pH superior a 8,40. Se sabe que un método de medición del pH para este tipo de soluciones proporciona valores que se distribuyen según una normal con media igual al nivel real de pH y desviación típica de 0,05. Supongamos que los niveles obtenidos en 10 mediciones independientes fueron los siguientes:

   8,30, 8,42, 8,44, 8,32, 8,43, 8,41, 8,42, 8,46, 8,37, 8,42

Aceptemos, además, que llevar a cabo el proceso con un reactivo que presente un nivel de pH menor o igual que 8,40 constituye un error muy serio.

(a) ¿Qué hipótesis nula se debe contrastar?

(b) ¿Cuál es la hipótesis alternativa?

(c) Al nivel de significación del 5%, ¿qué consejo se debería dar, usar o no usar la solución?

(d) ¿Cuál es el p valor del contraste?

5. En un anuncio de un dentífrico se mantiene que el uso del producto reduce significativamente el número de caries de los niños que están en la edad más propensa a sufrirlas. El número de caries al año para ese grupo de edad sigue una normal de media 3 y desviación típica 1. El chequeo de 2500 niños que usaron el dentífrico mostró que el número medio de caries por niño fue 2,95. Si se supone que la desviación típica del número de caries de los niños que usan el dentífrico continúa siendo 1:

   (a) Al nivel de significación del 5%, ¿estos datos son lo suficientemente fuertes para asegurar que lo dicho en el anuncio es correcto?

   (b) ¿Constituye esto una razón suficientemente significativa para que los niños utilicen este dentífrico?

6. Un granjero mantiene que puede producir tomates más grandes. Para contrastarlo se utiliza una variedad de tomate con un diámetro medio de 8,2 centímetros y desviación típica de 2,4 centímetros. Si la media de una muestra de 36 tomates de esta variedad, extraída de la producción del granjero, fue de 9,1 centímetros, ¿prueba esto que el tamaño medio es realmente mayor? Asuma que la desviación típica poblacional continúa siendo 2,4 y utilice un nivel de significación del 5%.

7. Supongamos que, tras el contraste descrito en el ejemplo 9.4, la compañía de tabaco está aún más convencida de que el contenido medio en nicotina de sus cigarrillos es menor que 1,5 miligramos por cigarrillo. ¿Debería sugerir otro contraste? ¿Con el mismo tamaño muestral?

8. Los datos siguientes provienen de una distribución normal con desviación típica 4.

   105, 108, 112, 121, 100, 105, 99, 107, 112, 122, 118, 105

   Utilícelos para contrastar la hipótesis nula de que la media de la población es menor o igual que 100:

   (a) Al nivel de significación del 5%.

   (b) Al nivel de significación del 1%.

   (c) ¿Cuál es el p valor?

9. Una compañía que produce un determinado refresco mantiene que sus máquinas dispensan, en media, 6 onzas por vaso, con una desviación típica de 0,14 onzas. Un consumidor se muestra escéptico al respecto, pues considera que la cantidad media servida es menor que 6 onzas. Para obtener información se selecciona una muestra de tamaño

100. Si la cantidad media por vaso fue de 5,6 onzas para esta muestra, ¿qué conclusiones se pueden extraer? Indique las hipótesis nula y alternativa y calcule el p valor resultante.

10. El contraste, a nivel de significación α, de

$$H_0: \mu = \mu_0 \quad \text{frente a} \quad H_1: \mu > \mu_0$$

es el mismo que para contrastar

$$H_0: \mu \leq \mu_0 \quad \text{frente a} \quad H_1: \mu > \mu_0$$

¿Parece esto razonable? ¡Explique por qué!

9.4 Contrastes de la *t* para la media de una población normal: caso de varianza desconocida

Previamente se ha asumido que el único parámetro desconocido de la distribución normal de la población era la media. Sin embargo, el caso más común es que también la desviación típica poblacional σ sea desconocida. En este apartado se mostrará cómo llevar a cabo contrastes de hipótesis referidos a la media en esta situación.

Para empezar, supongamos que podemos observar los resultados de una muestra aleatoria de tamaño n procedente de una población normal con media desconocida μ y desviación típica desconocida σ; y supongamos también que se desea utilizar los datos muestrales para contrastar la hipótesis nula

$$H_0: \mu = \mu_0$$

frente a la hipótesis alternativa

$$H_1: \mu \neq \mu_0$$

Como en el apartado anterior, parece razonable rechazar H_0 cuando el estimador puntual de la media poblacional μ –es decir, la media muestral $\overline{X}$ – esté muy separada de μ_0. Sin embargo, en la sección 9.3 se vio que la separación que justifica que se rechace H_0 depende de la desviación típica σ. En concreto, se demostró que el contraste, a nivel de significación α, rechazaba H_0 cuando $|\overline{X} - \mu_0|$ era como mínimo $z_{\alpha/2}\sigma/\sqrt{n}$ o, equivalentemente, cuando

$$\frac{\sqrt{n}\,|\overline{X} - \mu_0|}{\sigma} \geq z_{\alpha/2}$$

Ahora bien, si σ no es conocida, parece razonable estimarla por medio de la desviación típica muestral S, dada por

$$S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$$

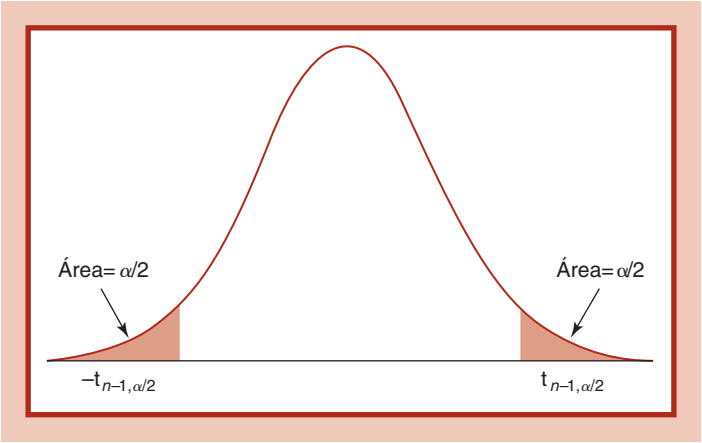


Figura 9.4 $P\{|T_{n-1}| \geq t_{n-1,\,\alpha/2}\} = \alpha$.

y hacer que el contraste actúe de forma que se rechace H_0 cuando el valor absoluto del estadístico del contraste T sea elevado, siendo

$$T = \sqrt{n}\,\frac{\overline{X} - \mu_0}{S}$$

Para determinar lo grande que debe ser $|T|$ para que esté justificado rechazar la hipótesis nula, a nivel de significación α, se necesita conocer su distribución cuando H_0 es cierta. Sin embargo, como se observó en la sección 8.6, cuando $\mu = \mu_0$, el estadístico T sigue una distribución t con $n - 1$ grados de libertad. Dado que el valor absoluto de esta variable aleatoria sobrepasa $t_{n-1,\,\alpha/2}$ con una probabilidad α (véase la figura 9.4), se sigue que el contraste, al nivel de significación α, de las hipótesis

$$H_0\colon \mu = \mu_0 \quad \text{frente a} \quad H_1\colon \mu \neq \mu_0$$

actuará, cuando σ sea desconocida, como sigue:

Rechazar H_0	si $\|T\| \geq t_{n-1,\,\alpha/2}$
No rechazar H_0	en otro caso

Este contraste, que está representado gráficamente en la figura 9.5, se denomina *contraste bilateral de la t*.

Si denotamos por v el valor del estadístico del contraste $T = \sqrt{n}(\overline{X} - \mu_0)/S$, el p valor asociado a los datos coincide con la probabilidad de que una variable aleatoria t con $n - 1$ grados de libertad, en valor absoluto, sea como mínimo $|v|$, la cual es igual al doble de la probabilidad de que una variable aleatoria t con $n - 1$ grados de libertad sea mayor o igual que $|v|$. (Esto es, el p valor es la probabilidad de que el estadístico del contraste tome un valor al menos tan grande como el valor absoluto del que se ha observado, si la hipótesis nula es cierta.) El contraste rechazará la hipótesis nula a cualquier nivel de significación mayor o igual que el p valor.

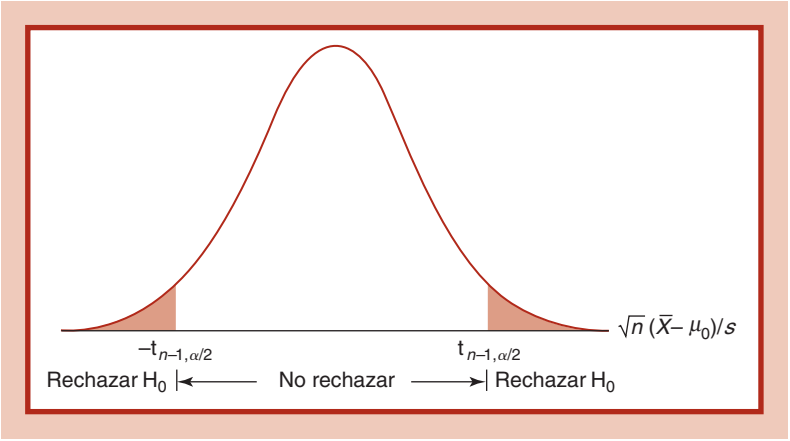


Figura 9.5 Contraste bilateral de la t, al nivel de significación α.

Si el valor del estadístico del contraste es v, se tiene que

$$p \text{ valor} = P\{|T_{n-1}| \geq |v|\}$$
$$= 2P\{T_{n-1} \geq |v|\}$$

donde T_{n-1} es una variable aleatoria t con $n-1$ grados de libertad.

Ejemplo 9.5 Entre los pacientes de una clínica que tienen altos niveles de colesterol, por encima de 240 mililitros por decilitro de suero sanguíneo, se han reclutado voluntarios para probar un nuevo medicamento reductor del colesterol en sangre. A un grupo de 60 voluntarios se les suministró el medicamento durante 60 días, y después se registraron las variaciones entre sus niveles de colesterol. Si la disminución media muestral fue de 6,8 con una desviación típica muestral de 12,1, ¿qué conclusiones se pueden sacar? Utilice un nivel de significación del 5%.

Solución Empecemos contrastando la hipótesis de que los cambios en los niveles de colesterol se deben simplemente al azar. Es decir, usemos los datos para contrastar las hipótesis

$$H_0: \mu = 0 \quad \text{frente a} \quad H_1: \mu \neq 0$$

donde μ representa la disminución media en los niveles de colesterol. El valor del estadístico del contraste T es

$$T = \frac{\sqrt{n}(\bar{X} - \mu_0)}{S} = \frac{\sqrt{40}(6{,}8)}{12{,}1} = 3{,}554$$

Puesto que, de la tabla D.2, $t_{39,\,0,025} = 2,02$, la hipótesis nula se debe rechazar, al nivel de significación del 5%. De hecho, el p valor resultante viene dado por

$$\begin{aligned} p \text{ valor} &= 2P\{T_{39} > 3,554\} \\ &= 0,0001 \quad \text{a partir del Programa 8-2} \end{aligned}$$

Así pues, a cualquier nivel de significación mayor que 0,0001, se rechazará la hipótesis de que el cambio en los niveles de colesterol es debido únicamente al azar.

Sin embargo, observe que pese a lo anterior no estaría justificado concluir que los cambios en los niveles de colesterol se deben específicamente al medicamento en cuestión y no a alguna otra razón. Por ejemplo, es bien conocido que cualquier medicación recibida por un paciente (tanto si la medicación es o no efectiva para la enfermedad que éste sufre) tiende a menudo a mejorar el estado del paciente (el *efecto placebo*). Otros factores adicionales podrían haber sido la causa de la reducción en los niveles sanguíneos de colesterol; por ejemplo, las condiciones mantenidas durante el periodo de prueba podrían haber afectado a los niveles de colesterol.

En realidad se tendría que concluir que el esquema de contraste anterior está diseñado de una forma muy pobre para conocer la efectividad del medicamento, puesto que para poder contrastar si un tratamiento particular es efectivo para una enfermedad concreta que podría estar afectada por distintas causas, se necesitaría diseñar un experimento que neutralice todas las restantes causas de cambio posibles, aparte del medicamento. El enfoque habitualmente aceptado para conseguir esto consiste en dividir aleatoriamente al conjunto de voluntarios en dos grupos: el primero recibe el medicamento, y al segundo (grupo de control) se le suministra un placebo (es decir, una pastilla con la misma apariencia y sabor que el medicamento pero que no tiene efecto fisiológico alguno). Los voluntarios no deberían conocer si pertenecen al grupo de tratamiento o al grupo de control. De hecho, esto es mejor que no lo conozcan ni siquiera los médicos (cuando así ocurre los experimentos se denominan *doblemente ciegos*), de forma que se pueda evitar que sus propios prejuicios y sesgos jueguen un papel en sus evaluaciones de los pacientes antes y después del experimento. Puesto que los dos grupos son seleccionados aleatoriamente dentro del conjunto de voluntarios se puede esperar que, en media, todos los factores que afectan a los dos grupos coinciden, excepto el derivado del hecho de haber recibido el medicamento o el placebo. Por consiguiente, cualquier diferencia que se encuentre entre los dos grupos puede atribuirse al medicamento. ■

El Programa 9-1 calcula el valor del estadístico del contraste T y el correspondiente p valor. Se puede utilizar tanto para llevar a cabo los contrastes unilaterales como los bilaterales. (Los contrastes unilaterales se presentarán a continuación.)

Ejemplo 9.6 Los datos históricos indican que el nivel de acidez media (pH) de la lluvia en una determinada región industrial de West Virginia es 5,2. Para contrastar si se ha producido recientemente algún cambio en este valor se midieron los niveles de acidez de 12 tormentas de lluvia del pasado año, y se obtuvieron los resultados siguientes:

6,1, 5,4, 4,8, 5,8, 6,6, 5,3, 6,1, 4,4, 3,9, 6,8, 6,5, 6,3

¿Son estos datos lo suficientemente fuertes para poder concluir, al nivel de significación del 5%, que la acidez de la lluvia ha cambiado con respecto a su valor histórico?

Solución Para contrastar la hipótesis de que no ha habido ningún cambio en la acidez, es decir, para contrastar

$$H_0: \mu = 5{,}2 \quad \text{frente a} \quad H_1: \mu \neq 5{,}2$$

se debe empezar calculando el valor del estadístico del contraste T. Ahora bien, puede fácilmente comprobarse que la media muestral y la desviación típica muestral son, respectivamente,

$$\overline{X} = 5{,}667 \quad \text{y} \quad S = 0{,}921$$

Por consiguiente, el valor del estadístico del contraste es

$$T = \sqrt{12}\,\frac{5{,}667 - 5{,}2}{0{,}921} = 1{,}76$$

Puesto que, de la tabla D.2 del Apéndice D, $t_{11,\,0{,}025} = 2{,}20$, la hipótesis nula no resulta rechazada, al nivel de significación del 5%. Esto es, los datos no son lo suficientemente fuertes para permitirnos concluir que la acidez de la lluvia ha cambiado, al nivel de significación del 5%.

Se podría haber resuelto este problema computando el p valor utilizando el Programa 9-1.

El valor de mu-cero es 5,2

El tamaño muestral es 12

Los valores de los datos son 6,1, 5,4, 4,8, 5,8, 6,6, 5,3, 6,1, 4,4, 3,9, 6,8, 6,5 y 6,3

El programa obtiene como valor del estadístico del contraste 1,755621

El p valor resultante es 0,1069365

Así pues, el p valor es 0,107 y, por tanto, la hipótesis nula no resulta rechazada ni siquiera al nivel de significación del 10%. ■

Supongamos que se desea contrastar la hipótesis nula

$$H_0: \mu \leq \mu_0$$

frente a la alternativa

$$H_1: \mu > \mu_0$$

En este caso se debería rechazar la hipótesis nula de que la media poblacional es menor o igual que μ_0 solamente cuando el estadístico del contraste

$$T = \sqrt{n}\,\frac{\overline{X} - \mu_0}{S}$$

sea significativamente grande (puesto que esto tiende a ocurrir cuando la media muestral es significativamente mayor que μ_0). Por consiguiente, el contraste, a nivel de significación α, actuará como sigue:

Rechazar H_0	si $T \geq t_{n-1,\,\alpha}$
No rechazar H_0	en otro caso

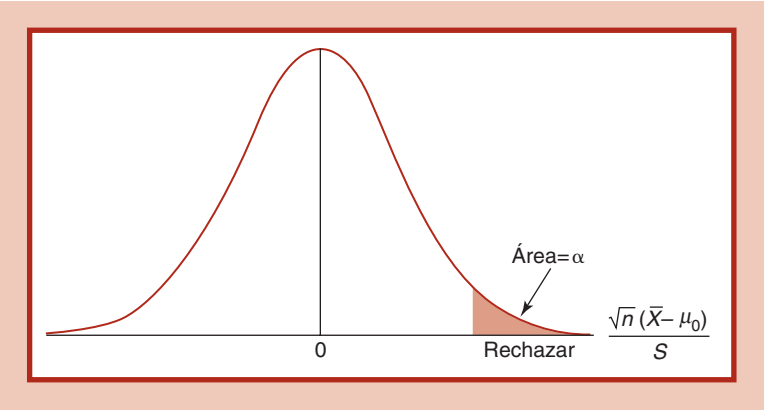


Figura 9.6 Contraste de H_0: $\mu \leq \mu_0$ frente a H_1: $\mu > \mu_0$.

La figura 9.6 muestra gráficamente la forma de actuar del contraste.

Equivalentemente, el contraste anterior se puede llevar a cabo obteniendo primero el valor, digamos ν, del estadístico del contraste T, y después calculando el p valor, que coincide con la probabilidad de que una variable aleatoria t con $n-1$ grados de libertad sea mayor o igual que ν. Esto es, si $T = \nu$,

$$p \text{ valor} = P\{T_{n-1} \geq \nu\}$$

Si se desea contrastar la hipótesis

$$H_0\text{: } \mu \geq \mu_0 \quad \text{frente a} \quad H_1\text{: } \mu < \mu_0$$

se actuará de forma similar. El contraste a nivel de significación α se basa de nuevo en el estadístico del contraste

$$T = \sqrt{n}\,\frac{\overline{X} - \mu_0}{S}$$

y actuará de la siguiente forma:

Rechazar H_0	si $T \leq -t_{n-1.\alpha}$
No rechazar H_0	en otro caso

Adicionalmente, el p valor es igual a la probabilidad de que una variable aleatoria t con $n-1$ grados de libertad sea menor o igual que el valor observado de T.

El Programa 9-1 permite obtener el valor del estadístico del contraste T y el p valor resultante. Si se dispone tan solo de los datos sumariales muestrales se puede utilizar el Programa 8-2 que computa las probabilidades relativas a las variables aleatorias t.

Ejemplo 9.7 El productor de una nueva cubierta de fibra de cristal para coches mantiene que la vida media de sus cubiertas es, como mínimo, de 50 000 millas. Para verificar esto se ha seleccionado una muestra aleatoria de 8 cubiertas que fueron, posteriormente, comprobadas por una agencia de consumo. Si la media muestral y la desviación típica muestral

La Estadística en perspectiva

¿Cuál es la hipótesis nula apropiada?

Supongamos que se sabe que los tubos de televisión producidos por una determinada compañía tienen una vida media de uso de 3000 horas. Un consultor externo mantiene que con un nuevo método de producción se consigue una mayor vida media. Para comprobar esto se diseña un programa piloto para producir una muestra de tubos mediante el nuevo método. ¿Cómo debería utilizar la compañía los datos resultantes?

A primera vista podría parecer que se deberían utilizar los datos para contrastar

$$H_0: \mu \leq 3000 \quad \text{frente a} \quad H_1: \mu > 3000$$

Si se hace esto, rechazar H_0 significaría una fuerte evidencia a favor de que el nuevo método de producción mejora la duración de los tubos de televisión. Sin embargo, el problema cuando se contrasta las anteriores hipótesis estriba en que, si el tamaño muestral es suficientemente grande, existe una posibilidad razonable de rechazar H_0 incluso en los casos en que la vida media muestral esté muy poco por encima de las 3000 horas; por ejemplo, si resulta ser 3001, de modo que no sea económicamente rentable llevar a cabo el cambio en el proceso de producción para conseguir un incremento tan reducido en la vida media. En realidad se deberían utilizar los datos para contrastar

$$H_0: \mu \leq 3000 + c$$

frente a

$$H_1: \mu > 3000 + c$$

donde c representa el mínimo aumento en la vida media que haga económicamente rentable llevar a cabo el cambio en el proceso de producción.

resultantes fueron, respectivamente, 47,2 y 3,1 (en miles de millas) contraste la tesis de la compañía.

Solución Para determinar si los anteriores datos están en concordancia con la hipótesis de que la vida media es, como mínimo, de 50 000 millas, se debería contrastar

$$H_0: \mu \geq 50 \quad \text{frente a} \quad H_1: \mu < 50$$

Rechazar la hipótesis H_0 estaría en contra de lo mantenido por el productor. El valor del estadístico del contraste T es

$$T = \sqrt{8}\,\frac{47{,}2 - 50}{3{,}1} = -2{,}55$$

Teniendo en cuenta que $t_{7,\,0{,}05} = 1{,}895$ y que se debe rechazar H_0 si T es menor o igual que $-t_{7,\alpha}$, se debe rechazar la hipótesis nula, al nivel de significación del 5%. Dado que $t_{7,\,0{,}01} = 2{,}998$, se debe también rechazar H_0 al nivel de significación del 1%. La ejecución del Programa 8-2 muestra que el p valor es igual a 0,019, lo cual indica que los datos manifiestan fuertemente que la tesis del productor no es correcta. ■

Se puede utilizar el contraste de la t incluso aunque la distribución subyacente no sea normal, siempre que el tamaño muestral sea razonablemente grande. Esto es cierto, por un lado, debido al teorema central del límite, que garantiza que la media muestral sigue aproximadamente una distribución normal sea cual sea la distribución poblacional; y, por otro lado, debido a que la desviación típica muestral será aproximadamente igual a σ. De hecho, puesto que, para valores grandes de n, la distribución t con $n-1$ grados de libertad es casi idéntica a la normal estándar, se verificará que $\sqrt{n}(\overline{X} - \mu_0)/S$ sigue aproximadamente una distribución normal estándar, si μ_0 es la media poblacional y el tamaño muestral n es suficientemente grande.

La tabla 9.2 incluye un resumen de todos los contrastes presentados en esta sección.

Problemas

1. Existe cierta variabilidad en las cantidades de fenobarbital contenidas en las cápsulas fabricadas por un determinado productor. Sin embargo, el productor mantiene que el valor medio por cápsula es de 2,0 miligramos. Para contrastar esto se seleccionó una muestra de 25 cápsulas cuyo valor medio fue de 19,7 miligramos, con una desviación típica muestral de 1,3. ¿Qué inferencia se debería concluir a partir de estos datos? En particular, ¿tienen los datos la fuerza suficiente para refutar la tesis del productor? Utilice un nivel de significación del 5%.
2. Un establecimiento de comida rápida tiene unas ventas medias de 2000 $ por día. Para contrastar si las cifras de negocio están cambiando debido al deterioro de la economía (que puede ser positivo o negativo para la industria de comida rápida), la dirección ha decidido registrar cuidadosamente las cifras de negocio de los 8 días próximos. Si los valores fueron

   2050, 2212, 1880, 2121, 2205, 2018, 1980, 2188

Tabla 9.2 Contrastes de hipótesis relativos a la media de una población normal con varianza desconocida σ^2.

$X_1, \ldots, X_n$ son los datos muestrales;

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \qquad S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$$

H_0	H_1	Estadístico del contraste TS	Contraste a nivel de significación α	p valor si TS $= v$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\sqrt{n}\dfrac{\overline{X} - \mu_0}{S}$	Rechazar H_0 si $\lvert TS \rvert \geq t_{n-1,\,\alpha/2}$ No rechazar H_0 en otro caso	$2P\{T_{n-1} \geq \lvert v \rvert\}$
$\mu \leq \mu_0$	$\mu > \mu_0$	$\sqrt{n}\dfrac{\overline{X} - \mu_0}{S}$	Rechazar H_0 si TS $\geq t_{n-1,\,\alpha}$ No rechazar H_0 en otro caso	$P\{T_{n-1} \geq v\}$
$\mu \geq \mu_0$	$\mu < \mu_0$	$\sqrt{n}\dfrac{\overline{X} - \mu_0}{S}$	Rechazar H_0 si TS $\leq -t_{n-1,\,\alpha}$ No rechazar H_0 en otro caso	$P\{T_{n-1} \leq v\}$

T_{n-1} representa una variable aleatoria t con $n-1$ grados de libertad; $t_{n-1,\,\alpha}$ y $t_{n-1,\,\alpha/2}$ son valores tales que $P\{T_{n-1} \geq t_{n-1,\,\alpha}\} = \alpha$ y $P\{T_{n-1} \geq t_{n-1,\,\alpha/2}\} = \alpha/2$.

Perspectiva histórica

(A. Barrington/Science Photo Library/Photo Researchers)

Ronald A. Fisher

En 1908, William Seeley Gosset, quien firmaba con el nombre de *Student*, publicó la distribución del estadístico de la t: $\sqrt{n}(\overline{X} - \mu)/S$. Fue un resultado importante, puesto que capacitó la realización de contrastes sobre las medias poblacionales cuando sólo se disponía de muestras de pequeño tamaño, como habitualmente ocurría en la destilería Guinness, donde Gosset estaba empleado. Sin embargo, en su época la importancia de ello fue prácticamente ignorada por la comunidad científica. Esto se debió principalmente al hecho de que sacar conclusiones a partir de muestras pequeñas estaba en contra de la opinión dominante en la época, en la que prevalecía el criterio de que "si el tamaño muestral es suficientemente grande, se debe sustituir σ por S y utilizar la distribución normal; y si el tamaño muestral no es suficientemente grande, no se debe utilizar la Estadística." Uno de los pocos que fueron conscientes de su importancia fue R. A. Fisher, quien en un artículo posterior corrigió ciertos aspectos técnicos del trabajo de Gosset. Pese a ello, el contraste de la t sólo fue reconocido y apreciado tras la publicación, en 1925, del libro de Fisher *Métodos estadísticos para investigadores*. Este libro tuvo un tremendo éxito, como demuestra el hecho de que se produjeran 11 ediciones en los siguientes 25 años. Pese a esta gran influencia, el libro, al igual que otros trabajos de Fisher, no es de fácil lectura. De hecho, un colega de Fisher llegó a decir que "ningún estudiante que no lo haya leído previamente debería intentar leerlo."

William S. Gosset

(Nota: Fotografía de Gosset incluida en "Student: Biografía estadística de William Seeley Gosset". Se basa en los escritos de E. S. Pearson, editada y aumentada por R. L. Plackett con la asistencia de G. A. Barnard. Clarendon Press, Oxford, 1990. Fotografía sacada de Annals of Eugenics, 1939, vol. 9.)

(a) ¿Cuáles son las hipótesis nula y alternativa?

(b) ¿Estos datos son lo suficientemente significativos para probar, al nivel del 5%, que se ha producido un cambio?

(c) ¿Qué ocurre al nivel de significación del 1%?

(d) Si puede utilizar el Programa 9-1 o algún *software* equivalente, obtenga el p valor.

3. Para contrastar la hipótesis de que una población normal tiene una media de 100 se ha seleccionado una muestra de tamaño 10. Si la media muestral es 110, ¿se debería rechazar la hipótesis nula si se conociera lo siguiente?

   (a) La desviación típica poblacional es igual a 15.

   (b) La desviación típica poblacional es desconocida y la desviación típica muestral resultó ser igual a 15.

   Utilice un nivel de significación del 5%.

4. El número de comidas servidas diariamente en la cafetería de una determinada escuela durante el último año se distribuye según una normal de media 300. En el año actual se ha decidido cambiar el menú y hacerlo más saludable, y la administración desea contrastar la hipótesis de que el número medio de comidas servidas no ha cambiado. Las comidas servidas en una muestra de 12 días se reflejan a continuación:

   312, 284, 281, 295, 306, 273, 264, 258, 301, 277, 280, 275

¿Se debe rechazar la hipótesis de que la media es igual a 300? Use un nivel de significación del:

(a) 10%

(b) 5%

(c) 1%

5. Un científico oceanográfico pretende contrastar si la profundidad media del mar en una determinada zona es de 55 brazas, tal como se había registrado previamente. Para ello tomó medidas en 36 puntos de la zona elegidos aleatoriamente, y obtuvo una media muestral de 56,4 brazas con una desviación típica de 5,1 brazas. ¿Son estos datos lo suficientemente significativos para rechazar la hipótesis nula de que la profundidad media es de 55 brazas? Al nivel de significación del:

   (a) 10%

   (b) 5%

   (c) 1%

6. Hace 20 años, los alumnos de primer curso de enseñanza media podían hacer, en media, 24 flexiones gimnásticas en 60 segundos. Para ver si esto continúa igual en la actualidad se ha seleccionado una muestra de 36 alumnos de primer curso. Si la media muestral resultó ser de 22,5 flexiones con una desviación típica muestral de 3,1, ¿se puede concluir que la media ha dejado de ser igual a 24? Utilice un nivel de significación del 5%.

7. El tiempo medio de respuesta a un estímulo es de 0,8 segundos para una determinada especie de cerdos. Veintiocho cerdos de esta especie recibieron 2 onzas de alcohol y, después, se midieron sus tiempos de reacción al estímulo citado. Si su tiempo medio de respuesta fue de 1,0 segundos con una desviación típica muestral de 0,3 segundos, ¿se puede concluir que el alcohol afecta al tiempo medio de respuesta? Utilice un nivel de significación del 5%.

8. Estudios previos demuestran que los ratones ganan, en media, 5 gramos de peso durante los diez primeros días que siguen al destete. Un grupo de 36 ratones recibieron un edulcorante artificial añadido a su comida. Su ganancia media en peso fue de 4,5 gramos, con una desviación típica de muestral de 0,9 gramos. ¿Se puede concluir, al nivel del 1%, que el edulcorante añadido ha tenido algún efecto?

9. Utilice los resultados de los partidos del último domingo de la Liga Nacional de Fútbol Profesional (NFL) de Estados Unidos para contrastar la hipótesis de que la puntuación media de los equipos vencedores es de 26,2 puntos. Utilice un nivel de significación del 5%.

10. Utilice los resultados de los partidos del último domingo de la liga de béisbol para contrastar que el número medio de carreras de los equipos vencedores es de 5,6. Utilice un nivel de significación del 5%.

11. Una panadería fue llevada a juicio por vender barras de pan con un peso inferior a las 24 onzas anunciadas. En su defensa, la panadería mantuvo que el peso anunciado de 24 onzas hacía referencia no al peso de cada barra particular, sino a que el peso medio calculado sobre el total de barras fabricadas. La acusación refutó lo indicando poniendo

como prueba que, en una muestra aleatoria de 20 barras, el peso medio muestral fue de 22,8 onzas, con una desviación típica muestral de 1,4 onzas. La jueza falló que el anuncio de las 24 onzas por barra se podía considerar aceptable si el peso medio muestral fuera como mínimo de 23 onzas.

(a) ¿Qué hipótesis estaba contrastando?

(b) Al nivel de significación del 5%, ¿qué debería fallar la jueza?

12. En un estudio recientemente publicado se mantenía que el salario medio anual de los profesores de las universidades de Estados Unidos era de 87 800 $. Los estudiantes de una universidad privada intuyen que el salario medio de sus profesores supera la cifra del estudio; en consecuencia, deciden contrastar las hipótesis

$$H_0: \mu \leq 87{,}800 \quad \text{frente a} \quad H_1: \mu > 87{,}800$$

siendo μ el salario medio de los profesores de su universidad. En una muestra aleatoria de 10 profesores, éstos declararon los siguientes salarios anuales (en unidades de 1000 dólares):

91,0, 79,8, 102,0, 93,5, 82,0, 88,6, 90,0, 98,6, 101,0, 84,0

(a) Al nivel de significación del 10%, ¿se ha de rechazar la hipótesis nula?

(b) ¿Qué ocurriría al nivel del 5%?

13. En un anuncio se indica que un determinado coche puede recorrer como mínimo 31 millas por galón de gasolina, conduciendo por autopista trayectos superiores a 100 millas. Supongamos que las millas recorridas por galón en 8 pruebas independientes (que consistieron en un recorrido sin paradas de 100 millas por autopista) fueron

28, 29, 31, 27, 30, 35, 25, 29

(a) Si se quisiera contrastar si los datos refutan lo que se mantiene en el anuncio, ¿qué hipótesis nula se debería tomar?

(b) ¿Cuál sería la hipótesis alternativa?

(c) Al nivel de significación del 5%?, ¿el anuncio resulta falso?

(d) ¿Qué ocurre al nivel del 1%?

14. Un productor mantiene que la duración media de las baterías que fabrica es como mínimo de 250 horas de uso. De una muestra de 20 baterías se observaron las duraciones siguientes:

237, 254, 255, 239, 244, 248, 252, 255, 233, 259, 236, 232, 243, 261, 255, 245, 248, 243, 238, 246

(a) Al nivel de significación del 5%, ¿están estos datos en concordancia con lo mantenido por el productor?

(b) ¿Qué ocurre al nivel del 1%?

15. Un oficial de la compañía de aguas insiste en que el consumo diario medio de agua por hogar en una cierta zona residencial es, cómo mínimo, de 400 galones. Para contrastar esto, se seleccionó una muestra aleatoria de 25 hogares de la zona. La media muestral de los consumos resultó ser 367, con una desviación típica muestral de 62. ¿Está en concordancia esto con lo que mantiene el oficial?

16. Una compañía suministra planchas de plástico para uso industrial. Se ha producido un nuevo tipo de plástico, y la compañía desearía demostrar a un asesor independiente que la resistencia media a la tensión de este nuevo producto es mayor que 30,0, medida en libras por pulgada cuadrada necesarias para romper la plancha. Con una muestra aleatoria de tamaño 12 se midieron las resistencias siguientes:

    30,1, 27,8, 32,2, 29,4, 24,8, 31,6, 28,8, 29,4, 30,5, 27,6, 33,9, 31,4

    (a) Al nivel de significación del 5% ¿prueban estos datos que la resistencia media a la tensión es mayor que 30 libras por pulgada al cuadrado?

    (b) ¿Cuál fue la hipótesis nula planteada en el apartado (a)?

    (c) Si la respuesta en (a) fue negativa, ¿establecen los datos que la resistencia media a la tensión es menor o igual que 30 libras por pulgada al cuadrado?

17. Una científica en medicina cree que la temperatura media basal de los individuos sanos ha aumentado a lo largo del tiempo y que es ahora superior a 98,6º F (37º C). Para contrastar esto, la científica ha seleccionado aleatoriamente a 100 individuos sanos. Si la temperatura media muestral resultó ser de 98,74º F, con una desviación típica muestral de 1,1º F, ¿prueban estos datos la intuición de la científica, al nivel de significación del 5%? ¿Y al nivel del 1%?

18. En 2001, los estudiantes que entraron en una determinada universidad habían obtenido una calificación media de 542 puntos en una prueba de acceso. Se ha seleccionado una muestra aleatoria de 20 estudiantes entre los que entraron en el año 2003, y sus calificaciones en la prueba de acceso habían sido las siguientes:

    542, 490, 582, 511, 515, 564, 500, 602, 488, 512, 518, 522, 505, 569, 575, 515, 520, 528, 533, 515

    ¿Prueban los datos dados que la calificación media en la prueba de acceso ha descendido por debajo de 542? Utilice un nivel de significación del 5%.

9.5 Contrastes de hipótesis sobre proporciones poblacionales

En esta sección se considerarán contrastes referidos a la proporción de individuos de una población que presentan una determinada característica. Se supondrá que la población es muy grande (en teoría, de tamaño infinito), y se denotará como p a la proporción de individuos de la población que presentan la característica citada. Nuestro interés se centrará en contrastar la hipótesis nula

$$H_0: p \leq p_0$$

frente a la alternativa

$$H_1: p > p_0$$

siendo p_0 un valor prefijado de antemano.

Si se extrae una muestra aleatoria de n individuos de la población, y se representa por X el número de elementos de la muestra que presentan la característica, se tendrá que X sigue una distribución binomial de parámetros n y p. Ahora bien, está claro que se debería rechazar la hipótesis nula de que la proporción poblacional es menor o igual que p_0 sólo cuando X sea suficientemente grande. Por consiguiente, si el valor observado de X es x, el p valor de los datos será igual a la probabilidad de que X tome un valor al menos tan grande como el observado, si se asume que p es igual a p_0 (que es el mayor valor posible bajo la hipótesis nula). Es decir, si x es el valor observado de X, se tiene que

$$p \text{ valor} = P\{X \geq x\}$$

donde X representa una variable aleatoria binomial de parámetros n y p_0.

El p valor se puede calcular mediante la aproximación normal o bien con el Programa 5-1, que computa las probabilidades binomiales. La hipótesis nula deberá rechazar a cualquier nivel de significación mayor o igual que el p valor.

Ejemplo 9.8 Un renombrado educador mantiene que más de la mitad de la población de Estados Unidos está preocupada por la falta de programas pedagógicos en la televisión. Para recoger datos sobre esta cuestión, un servicio nacional de encuestas ha seleccionado aleatoriamente y entrevistado a 920 individuos. Si 478 de ellos (un 52%) ha declarado estar preocupado por la falta de programas educativos en televisión, ¿prueban estos datos la tesis mantenida por el educador?

Solución Para que la tesis del educador se pueda probar, los datos se deberían mostrar suficientemente en contra de la hipótesis nula de que un 50%, como máximo, de la población está preocupado con el tema en cuestión. Así pues, se tendrían que utilizar los datos para contrastar

$$H_0: p \leq 0{,}50 \quad \text{frente a} \quad H_1: p > 0{,}50$$

Puesto que 478 individuos de la muestra declararon estar preocupados con el tema, se tiene que el p valor de estos datos es

$$\begin{aligned} p \text{ valor} &= P\{X \geq 478\} && \text{siendo } X \text{ una binomial } (920, 0{,}50) \\ &= 0{,}1243 && \text{obtenido a partir del Programa 5-1} \end{aligned}$$

Con un p valor tan elevado no se puede concluir que la tesis del educador resulte probada. Pese a que los datos se decantan claramente a favor de la tesis del educador, ya que el 52%

de los encuestados se mostraron preocupados por la ausencia de programas pedagógicos en televisión, este resultado tiene una razonable posibilidad de ocurrencia si la tesis del educador fuera incorrecta; por consiguiente, la hipótesis nula no resulta rechazada.

Si no se dispusiera del Programa 5-1 se podría aproximar el p valor mediante la aproximación de la normal a las probabilidades binomiales. Puesto que $np = 920(0{,}50) = 460$ y $np(1-p) = 460(0{,}5) = 230$, se obtendría lo siguiente:

$$\begin{aligned} p \text{ valor} &= P\{X \geq 478\} \\ &= P\{X \geq 477{,}5\} \quad \text{corrección por continuidad} \\ &= P\left\{\frac{X - 460}{\sqrt{230}} \geq \frac{477{,}5 - 460}{\sqrt{230}}\right\} \\ &\approx P\{Z \geq 1{,}154\} = 0{,}1242 \end{aligned}$$

Se ve que el p valor obtenido mediante la aproximación normal resulta estar bastante próximo al p valor exacto obtenido con el Programa 5-1. ■

A modo de ejemplo sobre otro tipo de situaciones en las que uno puede estar interesado en contrastar una hipótesis referida a un parámetro de la binomial, consideremos una cadena de producción cuyos artículos fabricados se clasifican como aceptables o defectuosos. Una hipótesis habitual es la de asumir que los artículos producidos resultan ser, independientemente, defectuosos con cierta probabilidad p; así pues, el número de artículos defectuosos en una muestra de tamaño n sigue una distribución binomial de parámetros n y p.

Ejemplo 9.9 Un productor de chips de ordenador mantiene que, como máximo, un 2% de los chips que produce son defectuosos. Una compañía electrónica, impresionada por la tesis mantenida, ha comprado una gran cantidad de chips. Para comprobar la tesis del productor, la compañía ha decidido examinar una muestra de 400 chips de los citados. Si 13 de ellos resultaron ser defectuosos (un 3,25%), al nivel de significación del 5%, ¿prueban estos datos que la tesis del productor es errónea?

Solución Si p representa la probabilidad de que un chip resulte defectuoso, se debería contrastar la hipótesis nula

$$H_0\text{: } p \leq 0{,}02 \quad \text{frente a} \quad H_1\text{: } p > 0{,}02$$

Es decir, para que los datos puedan probar que la tesis del productor es errónea, se tendría que establecer dicha tesis como hipótesis nula. Puesto que 13 de los 400 chips observados fueron defectuosos, el p valor será igual a la probabilidad de que el número de piezas defectuosas observadas sea mayor o igual que 13 si p fuera igual a 0,02 (el mayor valor posible bajo H_0). Por consiguiente,

$$\begin{aligned} p \text{ valor} &= P\{X \geq 13\} \quad \text{siendo } X \text{ una binomial } (400, 0{,}02) \\ &= 0{,}0619 \quad \text{obtenido a partir del Programa 5-1} \end{aligned}$$

de donde se desprende que, aunque los datos están claramente en contra de la tesis del productor, no son lo suficientemente fuertes para rechazarla, al nivel de significación del 5%.

Si se hubiera utilizado la aproximación normal se obtendría el siguiente p valor:

$$\begin{aligned} p \text{ valor} &= P\{X \geq 13\} && \text{siendo } X \text{ una binomial } (400, 0{,}02) \\ &= P\{X \geq 12{,}5\} && \text{corrección por continuidad} \\ &= P\left\{\frac{X-8}{\sqrt{8(0{,}98)}} \geq \frac{12{,}5-8}{\sqrt{8(0{,}98)}}\right\} \\ &\approx P\{Z \geq 1{,}607\} && \text{donde } Z \text{ representa una normal estándar} \\ &= 0{,}054 \end{aligned}$$

Así pues, aunque el p valor obtenido mediante la aproximación normal no está tan próximo al p valor real como sería deseable, tiene la suficiente precisión para conducirnos a la conclusión correcta, en el sentido de que los datos no son lo suficientemente fuertes para que se pueda rechazar la hipótesis nula, al nivel de significación del 5%. ■

De nuevo, denotemos por p la proporción de miembros de la población que presentan una determinada característica y supongamos que se desea contrastar

$$H_0: p \geq p_0$$

frente a

$$H_1: p < p_0$$

siendo p_0 un valor prefijado de antemano. Esto es, pretendemos contrastar la hipótesis nula de que la proporción de elementos de la población que presentan esa característica es, como mínimo, p_0 frente a la hipótesis alternativa de que dicha proporción es menor que p_0. Si se selecciona una muestra aleatoria de individuos de la población y resulta que x de ellos presentan la característica, el p valor correspondiente a los datos viene dado por

$$p \text{ valor} = P\{X \leq x\}$$

siendo X una variable aleatoria binomial de parámetros n y p_0.

Esto es, cuando la hipótesis nula es que p es mayor o igual que p_0, el p valor es igual a la probabilidad de poder obtener un valor menor o igual que el observado asumiendo que p es igual que p_0.

9.5.1 Contrastes bilaterales de *p*

El cálculo del p valor correspondiente a unos datos es ligeramente más complicado si se desea contrastar la hipótesis

$$H_0: p = p_0$$

frente a la alternativa bilateral

$$H_1: p \neq p_0$$

siendo p_0 un valor dado.

De nuevo, supongamos que se ha seleccionado una muestra aleatoria de tamaño n y que X denota el número de individuos de la muestra que presentan la característica de interés. Se deseará rechazar H_0 cuando X/n, la proporción de elementos de la muestra que presentan la característica, es mucho menor o bien mucho mayor que p_0; o, equivalentemente, cuando X es mucho menor o mucho mayor que np_0. Puesto que se desea que la probabilidad de rechazar la hipótesis nula sea menor o igual que α cuando p_0 es la verdadera proporción poblacional, se puede conseguir este objetivo cuando, asumiendo que H_0 es cierta, se rechaza esta hipótesis para valores grandes y para valores pequeños con una probabilidad de $\alpha/2$, en ambos casos. Es decir, se rechazará H_0 si se observa un valor x para el que se verifique o bien la probabilidad de que X sea mayor o igual que x es menor o igual que $\alpha/2$, o bien la probabilidad de que X sea menor o igual que x es menor o igual que $\alpha/2$.

Por consiguiente, si el valor observado de X es x, se rechazará H_0 bien si

$$P\{X \leq x\} \leq \frac{\alpha}{2}$$

o bien si

$$P\{X \geq x\} \leq \frac{\alpha}{2}$$

siendo X una variable aleatoria binomial de parámetros n y p_0. De aquí se desprende que el contraste, a nivel de significación α, rechazará H_0 si

$$\text{Min}\{P\{X \leq x\}, P\{X \geq x\}\} \leq \frac{\alpha}{2}$$

o, equivalentemente, si

$$2\,\text{Min}\{P\{X \leq x\}, P\{X \geq x\}\} \leq \alpha$$

donde X es una binomial (n, p_0). Así pues, si x de los n individuos de la muestra aleatoria extraída presentan la característica, el p valor del contraste de

$$H_0: p = p_0 \quad \text{frente a} \quad H_1: p \neq p_0$$

es el siguiente:

$$p \text{ valor} = 2\,\text{Min}\{P\{X \leq x\}, P\{X \geq x\}\}$$

siendo X una variable aleatoria binomial de parámetros n y p_0.

Puesto que habitualmente suele ser evidente cuál de las dos probabilidades que aparecen en la expresión anterior del p valor es la menor (si $x \leq np_0$, casi siempre será la pri-

mera; en caso contrario, será la segunda), por lo general, para obtener el p valor, tan solo será necesario utilizar en una ocasión el Programa 5-1 o bien la aproximación normal.

Ejemplo 9.10 Los datos históricos indican que el 4% de los componentes fabricados en una determinada cadena de producción resultan defectuosos. Acaba de concluir una áspera disputa laboral, y la dirección desea saber si esto ha ocasionado un cambio en la cifra del 4%. Si se extrae una muestra aleatoria de 500 componentes y 16 de ellos resultaron defectuosos (un 3,2%), al nivel de significación del 5%, ¿esto nos proporciona una evidencia suficiente para concluir que efectivamente se ha producido un cambio?

Solución Para que se pueda concluir que se ha producido dicho cambio, los datos habrán de ser lo suficientemente fuertes para poder rechazar la hipótesis nula si se contrasta

$$H_0: p = 0{,}04 \quad \text{frente a} \quad H_1: p \neq 0{,}04$$

donde p representa la probabilidad de que un componente sea defectuoso. El p valor correspondiente a los 16 componentes defectuosos observados en una muestra de 500 de ellos es

$$p \text{ valor} = 2 \text{ Min}\{P\{X \leq 16\}, P\{X \geq 16\}\}$$

siendo X una variable aleatoria binomial (500, 0,04). Dado que $500 \times 0{,}04 = 20$, se ve que

$$p \text{ valor} = 2P\{X \leq 16\}$$

Puesto que X tiene media 20 y desviación típica $\sqrt{20(0{,}96)} = 4{,}38$, está claro que el doble de la probabilidad de que X sea menor o igual que 16 –un valor inferior a la media en menos de una vez la desviación típica– no será lo suficientemente pequeña para que se pueda rechazar la hipótesis nula. De hecho, se puede comprobar que

$$p \text{ valor} = 2P\{X \leq 16\} = 0{,}432$$

por consiguiente, no existe la suficiente evidencia para que se pueda rechazar la hipótesis de que la probabilidad de producir un componente defectuoso no ha cambiado. ■

La tabla 9.3 resume la forma de actuar de los distintos contrastes relativos a la proporción poblacional p.

Tabla 9.3 Contrastes de hipótesis relativos a p, la proporción de individuos de una población que presentan cierta característica

El número de elementos que, en una muestra de tamaño n, presentan la característica se representa por X, y B es una variable aleatoria binomial de parámetros n y p_0.

H_0	H_1	Estadístico del contraste TS	p valor si TS = x
$p \leq p_0$	$p > p_0$	X	$P\{B \geq x\}$
$p \geq p_0$	$p < p_0$	X	$P\{B \leq x\}$
$p = p_0$	$p \neq p_0$	X	$2 \text{ Min}\{P\{B \leq x\}, P\{B \geq x\}\}$

Problemas

Para resolver los siguientes problemas utilice bien el Programa 5-1 u otro *software* equivalente, o bien la aproximación normal para computar las probabilidades binomiales que se precisen.

1. Se sabe que un medicamento estándar es efectivo en un 72% de los casos en los que se utiliza para tratar una determinada infección. Se ha desarrollado un nuevo medicamento y se ha comprobado que ha sido efectivo en 42 de los 50 casos tratados. ¿Estos datos proporcionan la suficiente evidencia para demostrar que el nuevo medicamento es más efectivo que el antiguo? Encuentre el p valor resultante.

2. Un economista piensa que, como mínimo, el 60% de los inmigrantes que han estado trabajando en los servicios sanitarios de Estados Unidos durante más de un año tienen la sensación de que han estado subempleados con respecto a su preparación. Supongamos que 294 individuos de una muestra de tamaño 450 (un 65,3%) creen que han estado subempleados. Al nivel de significación del 5%, ¿esto proporciona la evidencia suficiente para que se pueda probar que la idea del economista es correcta? ¿Qué ocurre al nivel de significación del 1%?

3. Los hurtos constituyen un serio problema para los comercios minoristas. En un departamento de unos grandes almacenes se averiguó que 1 de cada 14 personas que entraban en el departamento realizaba hurtos. Para ayudar a aliviar este problema, desde hace 3 meses se ha decidido aumentar el número de guardias de seguridad, y este hecho se hizo público a gran escala. Para valorar su efecto, los almacenes seleccionaron aleatoriamente a 300 clientes a los que se les hizo un seguimiento por cámara. Si 18 de estas 300 personas realizaron hurtos, ¿ello prueba que, al nivel de significación del 5%, la política seguida funciona?

4. Denotemos como p a la proporción de votantes de una determinada ciudad que está a favor de una reestructuración de su equipo de gobierno y consideremos el contraste de la hipótesis

$$H_0: p \geq 0{,}60 \quad \text{frente a} \quad H_1: p < 0{,}60$$

   Sobre una muestra de n votantes, x de ellos se mostraron a favor de la reestructuración indicada. Indique si el contraste, al nivel de significación α, rechazará H_0 en cada uno de los casos siguientes:

   (a) $n = 100$, $x = 50$, $\alpha = 0{,}10$

   (b) $n = 100$, $x = 50$, $\alpha = 0{,}05$

   (c) $n = 100$, $x = 50$, $\alpha = 0{,}01$

   (d) $n = 200$, $x = 100$, $\alpha = 0{,}01$

5. Una candidata política mantiene que más de un 50% de la población está a favor de su candidatura. Para demostrarlo ha encargado un estudio a una empresa de sondeos. Ésta ha seleccionado a una muestra aleatoria de individuos de la población, a los que se les preguntó si estaban a favor de la candidata citada.

Perspectiva histórica

El primer contraste de hipótesis publicado en el que se "probó" la existencia de Dios

Resulta curioso el hecho que se va a comentar a continuación: el primer artículo publicado en el que se intentó demostrar la existencia de Dios mediante un contraste estadístico de hipótesis. En 1710, John Arbuthnot publicó un artículo en la revista *Philosophical Transactions of the Royal Society* en el que se analizaban las cifras de nacimientos de varones y hembras en los 82 años comprendidos entre 1629 y 1710. Descubrió que en todos esos años habían nacido sistemáticamente más hombres que mujeres. Arbuthnot mantuvo que esto no podía deberse simplemente al azar porque, si en cada nacimiento las probabilidades de ser hombre o mujer fueran iguales (lo que equivale a que en cada año es igualmente probable que nazcan más varones o más hembras), la probabilidad de los resultados observados es $(1/2)^{82}$. A partir de esto, razonó que se debía rechazar la hipótesis de que los datos observados se debieran únicamente al azar [en nuestras palabras, el p valor del contraste de H_0: $p = 1/2$ frente a H_1: $p \neq 1/2$ es igual a $2(1/2)^{82}$]. Tras ello, Arbuthnot mantuvo que el resultado se debía atribuir a una planificación previa. Él creía que era beneficioso que nacieran más varones que hembras, ya que los primeros realizaban los trabajos más duros y, por ello, tendían a morir antes, por lo que concluyó que la planificación citada era obra de Dios. (Por razones hoy todavía no aclaradas, parece que la probabilidad de que un recién nacido sea varón es más próxima a 0,51 que a 0,50.)

(a) Para que se pueda probar la tesis de la candidata política, ¿cuáles deberína ser la hipótesis nula y la alternativa?

Considere los tres resultados siguientes y obtenga los p valores correspondientes.

(b) En una muestra de 100 votantes, 56 de ellos (un 56%) se mostraron a favor de su candidatura.

(c) Sobre una muestra de 200 votantes, 112 (un 56%) se mostraron a favor de su candidatura.

(d) En una muestra de 500 votantes, 280 (un 56%) estuvieron a favor.

Explique intuitivamente la discrepancia en los resultados, si es que hay alguna, pese a que en todos los casos (b), (c) y (d) las proporciones muestrales a favor de la candidata son idénticas.

6. El director de un renovado programa de noticias de televisión mantiene, frente a las empresas anunciantes, que al menos un 24% del total de televisores encendidos cuando se emite su programa le siguen. Esta cuota de pantalla del 24% es particularmente importante, puesto que los costes de los anuncios emitidos aumentan a partir de este nivel. Supongamos que 50 televisores de una muestra de 200 siguieron el programa.

(a) Al nivel de significación del 5%, ¿esto proporciona una evidencia suficiente para probar la tesis del director del programa de noticias?

(b) Al nivel de significación del 5%, ¿los datos proporcionan una evidencia suficiente para que probar que la tesis del responsable del programa es infundada?

(c) ¿Se puede manterner que los resultados muestrales proporcionan evidencia a favor o en contra de la tesis del director del programa?

(d) ¿Qué se debería hacer a continuación?

7. Tres compañías independientes están llevando a cabo una encuesta para determinar si más de la mitad de la población está a favor de una iniciativa para imponer limitaciones de tráfico en el centro de una determinada ciudad. Cada compañía de sondeos desea comprobar si existe evidencia de que más de la mitad de la población está a favor de esa iniciativa. Por consiguiente, las tres compañías citadas pretenden contrastar

$$H_0: p \leq 0{,}5 \quad \text{frente a} \quad H_1: p > 0{,}5$$

donde p representa la proporción de individuos de la población a favor de la iniciativa.

(a) Supongamos que la primera compañía de sondeos selecciona una muestra de 100 personas, de las que 56 se declaran a favor de la iniciativa. ¿Proporciona esto suficiente evidencia para que, al nivel de significación del 5%, se pueda rechazar la hipótesis nula y quede probado, por tanto, que más de la mitad de la población está a favor de la iniciativa en cuestión?

(b) Si la segunda compañía parte de una muestra de 120 individuos, de los cuales 68 dicen estar a favor de la iniciativa, ¿proporciona esto suficiente evidencia para que, al nivel de significación del 5%, se pueda rechazar la hipótesis nula?

(c) Si la tercera compañía utiliza una muestra de 110 individuos, de los cuales 62 están a favor de la iniciativa, ¿proporciona esto evidencia bastante para que, al nivel de significación del 5%, se pueda rechazar la hipótesis nula?

(d) Supongamos que las tres compañías juntan sus muestras para conseguir una muestra de 330 individuos, de los cuales 186 de ellos están a favor de la iniciativa. ¿Se consigue así tener suficiente evidencia para que, al nivel de significación del 5%, se rechace la hipótesis nula?

8. Un servicio de ambulancias mantiene que, como mínimo, en un 45% de las llamadas que atiende existe un peligro de muerte para el paciente. Para comprobar esto se selecciona una muestra de 200 llamadas del fichero de servicios realizados. Si en 70 de éstas efectivamente se atendieron emergencias con peligro de muerte, ¿resulta creíble la tesis del servicio de ambulancias? Use un nivel de significación del:

(a) 5%

(b) 1%

9. Un comercio ha recibido un envío de artículos de cierto tipo. Si se puede establecer que más de un 4% de los artículos recibidos son defectuosos, se devuelve el envío. Supongamos que en una muestra de 90 artículos se encontró que 5 de ellos eran defectuosos. ¿Se debería devolver el envío al proveedor? Utilice un nivel de significación del 10%. ¿Qué ocurriría al nivel del 5%?

10. Los editores de un periódico universitario mantienen que al menos un 75% de los estudiantes están a favor de las calificaciones tradicionales numéricas en lugar de aquéllas

en las que simplemente se indican aprobado/suspenso. Para obtener información al respecto, un decano selecciona aleatoriamente a una muestra de 50 estudiantes y observa que 32 de ellos se declaran a favor de las calificaciones tradicionales. ¿Estos datos concuerdan con la tesis de los editores del periódico? Utilice un nivel de significación del 5%.

11. En una reciente encuesta publicada por el Instituto de Investigación sobre Educación Superior se mantiene que un 22% de los nuevos estudiantes universitarios se califican a sí mismos como políticamente liberales. Si 65 de los 264 nuevos estudiantes de una muestra extraída en la Universidad de Berkeley, en California, declararon ser liberales, ¿se puede mantener, al nivel de significación del 5%, que el porcentaje en la Universidad de Berkeley es mayor que la cifra nacional?

12. Ha sido de conocimiento común durante algún tiempo que el 22% de la población tiene un arma de fuego en casa. En una encuesta realizada recientemente se encontró que 54 individuos de una muestra aleatoria de tamaño 200 tenían armas de fuego en sus domicilios. Al nivel de significación del 5%, ¿estos datos proporcionan la evidencia suficiente en contra de la cifra de uso común?

13. El tiempo medio que está encendida la luz roja de un semáforo es de 30 segundos. Por ello, un determinado individuo cree que tiene suerte si ha de esperar menos de 15 segundos cuando encuentra el semáforo en rojo. Este individuo asume que la probabilidad de que tenga suerte es de 0,5. Para contrastar esta hipótesis mide sus tiempos de espera en 30 semáforos. Si no ha tenido que esperar más de 15 segundos en 19 de ellos, ¿debería rechazar la hipótesis de que p es igual a 0,5?

    (a) Utilice un nivel de significación del 10%.

    (b) Use un nivel de significación del 5%.

    (c) ¿Cuál es el p valor resultante?

14. Un estudiante de estadística desea contrastar la hipótesis de que los sucesos de obtener cara o cruz, al lanzar una determinada moneda, son igualmente probables. El estudiante realiza 200 lanzamientos, y como resultado obtiene 116 caras y 84 cruces.

    (a) Al nivel de significación del 5%, ¿qué debe concluir el estudiante?

    (b) ¿Cuáles son las hipótesis nula y alternativa?

    (c) ¿Cuál es el p valor resultante?

15. Un veinticinco por ciento de las mujeres embarazadas fuma. Un científico desea contrastar la hipótesis de que ésta es también la proporción de fumadores en la población de mujeres que sufren embarazos ectópicos. Para ello, el científico selecciona una muestra de 120 mujeres que sufrieron recientemente embarazos ectópicos. Si 48 de esas mujeres resultaron ser fumadoras, ¿cuál es el p valor al contrastar las hipótesis

$$H_0: p = 0{,}25 \quad \text{frente a} \quad H_1: p \neq 0{,}25$$

siendo p la proporción de fumadores en la población de mujeres que han padecido embarazos ectópicos?

Términos clave

Hipótesis estadística: Una sentencia sobre la naturaleza de una población. A menudo la sentencia se enuncia en términos de un parámetro poblacional.

Hipótesis nula: Una hipótesis estadística que se desea contrastar.

Hipótesis alternativa: La alternativa a la hipótesis nula.

Estadístico del contraste: Una función de los datos muestrales. Dependiendo de su valor, la hipótesis nula será o no rechazada.

Región crítica: Si el valor del estadístico del contraste cae dentro de esta región, se debe rechazar la hipótesis nula.

Nivel de significación: Un pequeño valor fijado antes de llevar a cabo el contraste. Representa la máxima probabilidad de rechazar la hipótesis nula cuando ésta es cierta.

Contraste de la *Z*: Un contraste en el que la hipótesis nula establece que la media de una población normal con varianza conocida es igual a un determinado valor.

***p* valor:** El menor nivel de significación al cual se debe rechazar la hipótesis nula.

Contrastes unilaterales: Contrastes estadísticos de hipótesis en los que tanto la hipótesis nula como la alternativa enuncian que un determinado parámetro de la población es menor o igual que (o mayor o igual que) un valor dado.

Contraste de la *t*: Un contraste en el que la hipótesis nula establece que la media de una población normal con varianza desconocida es igual a un determinado valor.

Resumen

Una *hipótesis estadística* es una sentencia sobre los parámetros de una distribución poblacional.

La hipótesis que se debe contrastar se denomina *hipótesis nula* y se denota por H_0. La hipótesis alternativa se denota por H_1.

Un contraste de hipótesis se lleva a cabo a través de un estadístico del contraste, que es una función en los datos muestrales, y de una *región crítica*. Se rechaza la hipótesis nula si el valor del estadístico del contraste cae dentro de la región crítica; en caso contrario, no se rechaza. La región crítica se elige de forma que la probabilidad de rechazar la hipótesis nula cuando ésta es cierta no sobrepase un valor α prefijado de antemano, llamado *nivel de significación* del contraste. Los niveles de significación habituales son 0,10, 0,05 y 0,01. El nivel de significación del 5% es el más común en la práctica, esto es, $\alpha = 0{,}05$.

Puesto que el nivel de significación se fija igual a un valor pequeño existe solamente una probabilidad pequeña de rechazar H_0 si es cierta. Así pues, lo que se intenta hacer con un contraste estadístico de hipótesis es determinar si los datos están en concordancia con una hipótesis nula dada. Por consiguiente, rechazar H_0 tiene un alto poder probatorio en el sentido de que la hipótesis nula no parece ser consistente con los datos, mientras que el poder probatorio de no rechazar H_0 es mínimo en el sentido de que H_0 sea consistente con

los datos. Por esta razón, la hipótesis que se desee "probar estadísticamente" debería coincidir con la hipótesis alternativa, y quedaría probada si se rechaza la hipótesis nula.

En la práctica hay ocasiones en las que el nivel de significación no se fija de antemano, sino que, por el contrario, se lleva a cabo el contraste determinando el mínimo de los niveles de significación bajo los cuales se rechaza la hipótesis nula. Este nivel de significación mínimo se denomina *p* valor. Una vez determinado el *p* valor, la hipótesis nula se rechazará con cualquier nivel de significación mayor o igual que el *p* valor. Las siguientes reglas empíricas sintetizan grosso modo la posible utilidad del *p* valor:

p valor $>0{,}1$	Los datos proporcionan una débil evidencia en contra de H_0.
p valor $\approx 0{,}05$	Los datos proporcionan una moderada evidencia en contra de H_0.
p valor $< 0{,}05$	Los datos proporcionan una fuerte evidencia en contra de H_0.

---

1. *Contrastando* H_0: $\mu = \mu_0$ *frente a* H_1: $\mu \neq \mu_0$ *en una población normal con desviación típica* σ *conocida*: El contraste, a nivel de significación α, se basa en el estadístico del contraste

$$\sqrt{n}\,\frac{\overline{X} - \mu_0}{\sigma}$$

y actúa como sigue:

$$\begin{array}{ll} \text{Rechazar } H_0 & \text{si } \sqrt{n}\,\dfrac{|\overline{X} - \mu_0|}{\sigma} \geq z_{\alpha/2} \\ \text{No rechazar } H_0 & \text{en otro caso} \end{array}$$

Si el valor observado del estadístico del contraste es v, el *p* valor viene dado por

$$\begin{aligned} p \text{ valor} &= P\{|Z| \geq |v|\} \\ &= 2P\{Z \geq |v|\} \end{aligned}$$

donde Z representa una variable aleatoria normal estándar.

---

2. *Contrastando*

(1) H_0: $\mu \leq \mu_0$ *frente a* H_1: $\mu > \mu_0$

o

(2) H_0: $\mu \geq \mu$ *frente a* H_1: $\mu < \mu_0$

en una población normal con desviación típica σ *conocida*: Estos contrastes se denominan contrastes unilaterales. El contraste, a nivel de significación α, se basa en ambos casos en el estadístico del contraste $\sqrt{n}(\overline{X} - \mu_0)/\sigma$. El contraste en la situación (1) consiste en

$$\begin{array}{ll} \text{Rechazar } H_0 & \text{si } \sqrt{n}\,\dfrac{(\overline{X} - \mu_0)}{\sigma} \geq z_\alpha \\ \text{No rechazar } H_0 & \text{en otro caso} \end{array}$$

Alternativamente se puede llevar a cabo el contraste (1) calculando primero el p valor correspondiente a los datos observados. Si el valor del estadístico del contraste es v, el p valor coincide con

$$p \text{ valor} = P\{Z \geq v\}$$

siendo Z una variable aleatoria normal estándar. La hipótesis nula se rechazará con cualquier nivel de significación mayor o igual que el p valor resultante.

En la situación (2), a nivel de significación α, el contraste consiste en

$$\begin{array}{ll} \text{Rechazar } H_0 & \text{si } \sqrt{n}\,\dfrac{(\overline{X} - \mu_0)}{\sigma} \leq -z_\alpha \\ \text{No rechazar } H_0 & \text{en otro caso} \end{array}$$

Alternativamente si el valor del estadístico del contraste $\sqrt{n}(\overline{X} - \mu_0)/\sigma$ es v, el p valor viene dado por

$$p \text{ valor} = P\{Z \leq v\}$$

siendo Z una variable aleatoria normal estándar.

---

3. *Contrastes bilaterales de la t*

$$H_0: \mu = \mu_0 \quad \textit{frente a} \quad H_1: \mu \neq \mu_0$$

en una población normal cuya varianza es desconocida: Este contraste se basa en el estadístico del contraste

$$T = \frac{\sqrt{n}(\overline{X} - \mu_0)}{S}$$

donde n es el tamaño muestral y S es la desviación típica muestral. El contraste, a nivel de significación α, consiste en

$$\begin{array}{ll} \text{Rechazar } H_0 & \text{si } |T| \geq t_{n-1,\,\alpha/2} \\ \text{No rechazar } H_0 & \text{en otro caso} \end{array}$$

El valor $t_{n-1,\,\alpha/2}$ es aquel que verifica

$$P\{T_{n-1} > t_{n-1,\,\alpha/2}\} = \frac{\alpha}{2}$$

donde T_{n-1} es una variable aleatoria t con $n - 1$ grados de libertad. Este contraste se conoce con el nombre de *contraste de la t*.

Como alternativa, se puede llevar a cabo el contraste de la t si se calcula primero el valor del estadístico del contraste T. Si éste es igual a v, el p valor viene dado por

$$\begin{aligned} p \text{ valor} &= P\{|T_{n-1}| \geq |v|\} \\ &= 2P\{T_{n-1} \geq |v|\} \end{aligned}$$

siendo T_{n-1} una variable aleatoria t con $n - 1$ grados de libertad.

4. *Los contrastes de la t unilaterales*

(1) H_0: $\mu \leq \mu_0$ frente a H_1: $\mu > \mu_0$

o

(2) H_0: $\mu \geq \mu_0$ frente a H_1: $\mu < \mu_0$

en una población normal con varianza desconocida: Estos contrastes se basan de nuevo en el estadístico del contraste

$$T = \sqrt{n}\frac{\overline{X} - \mu_0}{S}$$

donde n es el tamaño muestral y S es la desviación típica muestral.
El contraste, a nivel de significación α, para (1) consiste en

Rechazar H_0	si $T \geq t_{n-1,\,\alpha}$
No rechazar H_0	en otro caso

Como alternativa se puede calcular el p valor. Si el valor del estadístico del contraste es ν, el p valor se calcula mediante

$$p \text{ valor} = P\{T_{n-1} \geq \nu\}$$

siendo T_{n-1} representa una variable aleatoria t con $n - 1$ grados de libertad.

A nivel de significación α, el contraste de (2) consiste en

Rechazar H_0	si $T \leq -t_{n-1,\,\alpha}$
No rechazar H_0	en otro caso

Si el valor de T es ν, el p valor del contraste (2) coincide con

$$p \text{ valor} = P\{T_{n-1} \leq \nu\}$$

5. *Contrastes de hipótesis relativos a proporciones*: Si p es la proporción de individuos de una población grande que presentan cierta característica, para contrastar

$$H_0\text{: } p \leq p_0 \quad \text{frente a} \quad H_1\text{: } p > p_0$$

se debe extraer una muestra aleatoria de n elementos de la población. El estadístico del contraste es X, el número de individuos de la muestra que presentan la característica. Si el valor de X es x, el p valor viene dado por

$$p \text{ valor} = P\{B \geq x\}$$

donde B es una variable aleatoria binomial de parámetros n y p_0.

Supongamos que se desea contrastar

$$H_0\text{: } p \geq p_0 \quad \text{frente a} \quad H_1\text{: } p < p_0$$

Si el valor observado del estadístico del contraste es x, el p valor viene dado por

$$p \text{ valor} = P\{B \leq x\}$$

donde, de nuevo, B representa una variable aleatoria binomial de parámetros n y p_0.

Las probabilidades de la binomial se pueden calcular con el Programa 5-1, o se pueden aproximar teniendo en cuenta la aproximación de la normal a la binomial.

Supongamos ahora que el contraste que se desea llevar a cabo es bilateral; esto es, se pretende contrastar

$$H_0: p = p_0 \quad \text{frente a} \quad H_1: p \neq p_0$$

Si el número de elementos de la muestra que presentan la característica es x, el p valor coincide con

$$p \text{ valor} = 2 \text{ Min}\{P\{B \leq x\}, P\{B \geq x\}\}$$

siendo B una variable aleatoria binomial de parámetros n y p_0.

Problemas de repaso y estudios de caso propuestos

1. Supongamos que quisiéramos explicar a una persona sin estudios en Estadística el significado de que un contraste estadístico haya rechazado la hipótesis nula de que la media de una población es igual a 0; esto es, que se haya rechazado H_0: $\mu = 0$, al nivel de significación del 5%. ¿Cuál de las siguientes sentencias es más precisa?

   (a) Los datos evidencian que la media poblacional difiere significativamente de 0.

   (b) La evidencia de los datos es lo suficientemente significativa para indicar que la media poblacional difiere de 0.

   ¿Qué es equívoco en la sentencia menos precisa de las dos?

2. Supongamos que el p valor resultante en un contraste estadístico fue igual a 0,11.

   (a) Al nivel de significación del 5%, ¿se debería rechazar la hipótesis nula?

   (b) ¿Podría decirse que es evidente la bondad de la hipótesis nula? Explique brevemente la respuesta.

3. Supongamos que en un periódico local se lee: "Un estudio reciente proporciona una significativa evidencia de que la media de las alturas de las mujeres ha aumentado a lo largo de los últimos veinte años."

   (a) ¿Se puede considerar que esta sentencia es precisa?

   (b) ¿Qué interpretación debería darse a esta sentencia?

4. Un hecho conocido desde hace tiempo, y aún hoy no explicado, es que durante los primeros años los gemelos suelen obtener menores calificaciones en los test de inteligencia y tienden a ser más lentos en adquirir habilidades de lenguaje que los no mellizos.

Recientemente, algunos psicólogos han especulado que esto se puede deber a que los padres dedican menos tiempo a los mellizos que a los hijos no mellizos. Esto se explicaría por el hecho de que los mellizos tienen que compartir la atención de los padres y, además, también influyen los motivos económicos, puesto que los mellizos representan un mayor coste para los padres y de ello se deriva un menor tiempo disponible de dedicación a sus hijos.

Proponga un estudio que se pudiera utilizar para contrastar la hipótesis de que los padres dedican menos tiempo a los hijos mellizos que a los no mellizos.

Si se asume que esta hipótesis es cierta, proponga un estudio que nos permita concluir que ésta es la causa del hecho conocido desde hace tiempo y aún no explicado.

5. La ruta que sigue actualmente un individuo para ir al trabajo le supone un tiempo medio de 40 minutos por viaje. Un amigo le sugiere una ruta alternativa que, según él, le hará perder menos tiempo. Supongamos que en 10 viajes, elegidos aleatoriamente siguiendo la ruta sugerida por su amigo, empleó los tiempos siguientes:

   44, 38,5, 37,5, 39, 38,2, 36, 42, 36,5, 36, 34

   ¿Estos datos evidencian que la nueva ruta es más corta en tiempo? Use un nivel de significación del:

   (a) 1%

   (b) 5%

   (c) 10%

6. Para contrastar la hipótesis nula de que

   $$H_0: \mu = 15 \quad \text{frente a} \quad H_1: \mu \neq 15$$

   se ha seleccionado una muestra de tamaño 12. Si la media muestral es de 14,4, calcule el *p* valor si se sabe que la desviación típica poblacional es igual a

   (a) 0,5

   (b) 1,0

   (c) 2,0

7. Se mantiene que más de un 30% de los estudiantes que entran en una determinada universidad presentan unos niveles de colesterol de, como mínimo, 200. Utilice los datos de los últimos 20 estudiantes de la lista del Apéndice A para contrastar esta hipótesis. Al nivel de significación del 5%, ¿qué conclusión se debe extraer?

8. Los psicólogos que se consideran a sí mismos discípulos de Alfred Adler creen que el orden de nacimiento tiene una gran influencia sobre la personalidad. Adler mantenía que los primogénitos (incluyendo los hijos únicos) tendían a tener mayor confianza en sí mismos, lo que potenciaba un mayor éxito en la vida que a sus hermanos. Por ejemplo, entre los primeros 102 miembros del Tribunal Supremo de Estados Unidos, un 55% de ellos fueron primogénitos, mientras que sobre el total de la población la proporción de primogénitos es sólo del 37%.

(a) Con estos datos del Tribunal Supremo, contraste la hipótesis de que la creencia de los seguidores de Adler es errónea y que, por tanto, el hecho de ser primogénito no influye sobre la personalidad.

(b) ¿Es el resultado de (a) una prueba convincente de la validez de la posición de los Adlerianos? (*Sugerencia*: Recuerde la *minería de datos.*)

(c) Proponga un estudio para intentar aprobar o desaprobar la teoría de Adler. Elija una muestra de personas exitosas (por ejemplo, a los 200 mejores jugadores de la liga de béisbol profesional) y trate de averiguar qué porcentaje de ellos son primogénitos.

9. Un individuo llamado Nicolás Caputo fue un empleado del condado de Essex, Nueva Jersey, durante un largo periodo. Una de sus funciones consistía en diseñar muestras para predecir si en las distintas elecciones ganaría el candidato demócrata o el republicano. Durante este tiempo los demócratas ganaron en 40 de las 41 ocasiones. Debido a esto, Caputo, que era demócrata, recibió el apodo de *el hombre del brazo de oro*. En 1985, los republicanos del condado de Essex demandaron judicialmente a Caputo, aduciendo que discriminaba en contra de ellos. Si usted fuera juez, ¿qué dictaminaría? ¡Explique por qué!

10. Una teoría reciente mantiene que la gente famosa tiene una mayor probabilidad de morirse en los seis meses que siguen a su cumpleaños que en los seis meses que lo preceden. Es decir, la tesis afirma que una persona famosa que haya nacido, por ejemplo, el 1 de julio tiene mayor probabilidad de morir entre el 1 de julio y el 31 de diciembre que entre el 1 de enero y el 1 de julio. El razonamiento estriba en que las personas famosas son más sensibles a la atención y al afecto que se les presta en su cumpleaños y esto refuerza su deseo de vivir en los meses anteriores. Una teoría contraria mantiene que la gente famosa tiene menos probabilidad de morirse en los 6 meses posteriores a su cumpleaños debido a la fortaleza que les aporta su fiesta de cumpleaños. Incluso hay muchas personas que mantienen que ambas teorías son falsas.

    Denotemos como p a la probabilidad de que una persona famosa muera a lo largo de los 6 meses siguientes a su fecha de cumpleaños, y consideremos el contraste

$$H_0\colon p = \frac{1}{2} \quad \text{frente a} \quad H_1\colon p \neq \frac{1}{2}$$

    (a) Supongamos que alguien ha recogido listas de 200 personas famosas en 25 áreas diferentes, y con ellas ha llevado a cabo 25 contrastes distintos dentro de cada área..

    Incluso aunque la hipótesis nula fuera cierta, ¿cuál es la probabilidad de que en al menos uno de los contrastes se rechace la hipótesis nula, al nivel de significación del 5%?

    (b) Intente recoger una lista de entre 100 y 200 personas famosas y utilice ésta para contrastar las hipótesis citadas.

11. Seleccione una muestra aleatoria de 16 mujeres de la lista incluida en el Apéndice A, y utilice los datos de sus pesos para contrastar la hipótesis nula de que el peso medio de todas las mujeres de la lista no es mayor que 110 libras. Utilice un nivel de significación del 5%.

12. Supongamos que los equipos A y B están jugando un partido de la Liga de Fútbol Nacional de Estados Unidos y que el equipo A va ganado por f puntos en la mitad del partido. Denotemos por S(A) y S(B) los puntos conseguidos en el partido por los equipos A y B, respectivamente, y sea $X = S(A) - S(B) - f$. Esto es, X denota la diferencia entre los puntos marcados en la segunda mitad del partido por los equipos A y B. Se mantiene que la distribución de X es normal con media 0 y desviación típica 14. Utilice los datos de varios partidos elegidos aleatoriamente para contrastar esa hipótesis.

13. El modelo de paseo aleatorio para los precios de una acción o de un bien asume que las diferencias sucesivas entre los logaritmos de los precios de cierre diarios constituyen una muestra aleatoria procedente de una población normal. Los datos siguientes muestran los precios de cierre del oro en 17 días laborables consecutivos de 1994. Utilice estos datos para contrastar la hipótesis de que la media de los cambios es igual a 0.

Precios de cierre					
387,10	391,00	389,50	391,00	395,00	396,25
388,00	391,95	390,25	390,50	393,50	395,45
389,65	391,05	388,00	394,00	396,25	

Observación: Los datos están ordenados por columnas. El primer valor es de 387,10; el segundo, de 388,00; el tercero, de 389,65; el cuarto, de 391,00, y así sucesivamente.

14. Una determinada hipótesis nula se debe rechazar cuando el valor del estadístico del contraste, TS, es grande. El valor observado del TS es 1,3. Supongamos que cuando la hipótesis nula es cierta, la probabilidad de que el TS sea mayor o igual que 1,3 es 0,063.

    (a) Al nivel de significación del 5%, ¿se debe rechazar la hipótesis nula?

    (b) ¿Y al nivel de significación del 10%?

    (c) ¿Cuál es el p valor resultante?

10

Contrastes de hipótesis relativas a dos poblaciones

Las estadísticas son como los terapeutas, ellos dan testimonio por las dos partes.

Fiorello La Guardia, anterior alcalde de la ciudad de Nueva York

Los números no mienten; y ellos no perdonan.

Harry Angstrom en *El conejo es rico* de John Updike

Se analiza la importancia de usar un control a la hora de contrastar una nueva medicina o un nuevo procedimiento, y se verá que esto a menudo se traduce en comparaciones entre parámetros de dos poblaciones diferentes. Se muestra cómo contrastar que dos poblaciones normales tienen la misma media poblacional, tanto si las varianzas poblacionales son conocidas como si son desconocidas. Se muestra cómo contrastar la igualdad de dos proporciones poblacionales.

10.1 Introducción

En un debate en curso de gran importancia surge si las megadosis –del orden de 25 000 a 30 000 miligramos diarios– de vitamina C pueden ser efectivas en el tratamiento de pacien-

tes que sufren de tumores cancerígenos. Por un lado, en la controversia se encontraba el gran químico norteamericano Linus Pauling, fuerte defensor de la terapia con vitamina C, al igual que un creciente número de investigadores; por otro lado estaban la gran mayoría de los terapeutas del cáncer. Aunque se habían diseñado un gran número de experimentos para contrastar si la vitamina C era terapéuticamente efectiva, existía una fuerte controversia con respecto a muchos de ellos. Algunos de estos experimentos registraban resultados negativos, y los partidarios de la vitamina C los cuestionaban basándose en que se utilizaban pequeñas dosis de vitamina. Una parte de la comunidad científica recibió con escepticismo otros experimentos realizados por el profesor Pauling y sus asociados. Para cerrar la contienda, en los años siguientes en la Clínica Mayo se planeó y se realizó un experimento definitivo. En ese famoso estudio, durante tres meses a un grupo de pacientes terminales de cáncer les suministraron grandes dosis de vitamina C, además de la medicación regular. Los demás pacientes recibieron un placebo junto con la medicación regular. Tras el periodo de tres meses, el experimento se cerró y se cortó el suministro de la vitamina C; se llevó a cabo un seguimiento de por vida de los pacientes para determinar si los pacientes que habían recibido la vitamina C tenían una duración de vida superior a la del grupo de control. Al final del experimento se emitió un resultado clasificatorio, que fue ampliamente difundido por los medios de comunicación, y mantenía que no existían diferencias significativas entre los dos grupos de pacientes. Muchos miembros de la comunidad médica utilizaron este experimento para desacreditar fuertemente el uso de la vitamina C como tratamiento terapéutico del cáncer; sin embargo, Pauling lo cuestionó y mantuvo que era irrelevante con respecto a la tesis mantenida por los partidarios de la vitamina C. Según la teoría de Pauling y otros, la vitamina C, mientras se tomaba, tenía un valor protector que desaparecía si se cortaba el tratamiento. Verdaderamente, de acuerdo con anteriores escritos de Pauling, una parada inmediata en el suministro de la vitamina C (tal como se había hecho en el estudio) en lugar de una parada gradual podía tener potenciales efectos negativos. La controversia continúa.

Es importante observar que para la Clínica Mayo no hubiera bastado con dar una megadosis de vitamina C a todos los pacientes voluntarios. Pues, aunque hubiera habido incrementos significativos en las duraciones de vida de estos pacientes en comparación con la distribución conocida de las duraciones de los enfermos de cáncer, no habría sido posible atribuir como causa de este incremento a la vitamina C. Por dos razones: i) no se habría eliminado el efecto placebo, según el cual cualquier tipo de tratamiento "extra" da un deseo adicional al paciente, y éste por sí mismo puede tener efectos beneficiosos; ii) el aumento de vida adicional podría deberse a factores totalmente ajenos al experimento. Así pues, para poder sacar una conclusión válida del experimento, era necesario tener un segundo grupo de pacientes voluntarios tratados exactamente de la misma manera que los primeros, pero que no recibieran la vitamina C sino una medicación que se pareciera y tuviera el mismo sabor que ésta. (Naturalmente, para asegurar que los dos grupos eran lo más similares posible, con excepción del suministro de la vitamina C, el conjunto de pacientes voluntarios fue dividido en dos grupos: elgrupo de tratamiento, cuyos miembros recibieron la vitamina C, y el grupo de control, cuyos miembros recibieron el placebo.) Así se dispuso de dos muestras separadas, y los datos resultantes con cada una de ellas se utilizaron para contrastar la hipótesis de que las medias de las duraciones de vida de estos dos grupos eran idénticas.

Realmente, en todas las situaciones en las que uno intenta estudiar el efecto de un determinado factor, tal como la administración de la vitamina C, se debe intentar mantener

todos los restantes factores constantes, de forma que cualquier cambio en los comportamientos pueda ser atribuido solamente al factor bajo estudio. Sin embargo, puesto que a menudo es difícil de conseguir fuera de los experimentos en las ciencias físicas, habitualmente es necesario considerar dos muestras –una que recibe el factor bajo estudio; y otra, el grupo de control que no recibe el factor– y luego determinar si existe una diferencia estadísticamente significativa en las respuestas de esas dos muestras. Por esta razón, los contrastes relativos a dos poblaciones muestreadas son muy importantes en una gran variedad de aplicaciones.

En este capítulo se mostrará cómo contrastar la hipótesis de que dos medias poblacionales son iguales cuando se dispone de una muestra procedente de cada población. En la sección 10.2 se supondrá que las distribuciones poblacionales subyacentes son normales, con varianzas conocidas. Aunque la hipótesis de varianzas poblacionales conocidas es inusual, el análisis presentado en este apartado será útil para mostrar cómo tratar otros casos más importantes en los que no se mantiene dicha hipótesis. De hecho, en la sección 10.3 se muestra cómo contrastar la hipótesis de que las medias de dos poblaciones son iguales cuando las varianzas son desconocidas, en el supuesto de que los tamaños muestrales sean grandes. El caso de que los tamaños muestrales no son grandes se trata en la sección 10.4. Para poder contrastar la hipótesis en esta situación resulta necesario asumir que las varianzas poblacionales desconocidas son iguales.

En la sección 10.5 se consideran situaciones en las que las dos muestras se relacionan por un natural emparejamiento entre los elementos de los dos conjuntos de datos. Por ejemplo, uno de los datos en la primera muestra puede referirse a la presión sanguínea de de un individuo antes de recibir la medicación, mientras que el dato correspondiente en la segunda muestra registra la presión sanguínea de esa misma persona tras recibir la medicación.

En la sección 10.6 se consideran contrastes de igualdad de dos proporciones binomiales.

10.2 Contraste de igualdad de medias de dos poblaciones normales: caso de varianzas conocidas

Supongamos que $X_1,\ldots,X_n$ es una muestra procedente de una población normal con media μ_x y varianza σ_x^2, y que $Y_1,\ldots,Y_m$ es una muestra independiente de la anterior procedente de una población normal con media μ_y y varianza σ_y^2. Asumiendo que se conocen las varianzas poblacionales σ_x^2 y σ_y^2, consideremos el contraste cuya hipótesis nula es que las dos medias poblacionales son iguales; esto es, consideremos el contraste de

$$H_0: \mu_x = \mu_y$$

frente a la hipótesis alternativa

$$H_1: \mu_x \neq \mu_y$$

Puesto que los estimadores de μ_x y μ_y son las medias muestrales

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \qquad \text{y} \qquad \overline{Y} = \frac{\sum_{i=1}^{m} Y_i}{m}$$

parece razonable que H_0 debería ser rechazada cuando $\overline{X}$ e $\overline{Y}$ estén muy separadas. Es decir, para una apropiada constante c, es razonable que la actuación frente al contrate sea

$$\begin{array}{ll} \text{Rechazar } H_0 & \text{si } |\overline{X} - \overline{Y}| \geq c \\ \text{Aceptar } H_0 & \text{en otro caso} \end{array}$$

Para determinar el valor de c, para un contraste a nivel de significación α, necesitamos primero obtener la distribución de $\overline{X} - \overline{Y}$. Ahora bien, $\overline{X}$ sigue una normal con media μ_x y varianza σ_x^2/n; e $\overline{Y}$, similarmente, sigue una normal con media μ_y y varianza σ_y^2/m. Puesto que la diferencia de variables aleatorias normales e independientes continúa estando normalmente distribuida, se sigue que $\overline{X} - \overline{Y}$ es normal con media

$$E[\overline{X} - \overline{Y}] = E[\overline{X}] - E[\overline{X}] = \mu_x - \mu_y$$

y varianza

$$\begin{aligned} \text{Var}(\overline{X} - \overline{Y}) &= \text{Var}(\overline{X}) + \text{Var}(-\overline{Y}) \\ &= \text{Var}(\overline{X}) + (-1)^2\text{Var}(\overline{Y}) \\ &= \text{Var}(\overline{X} + \text{Var}(\overline{Y}) \\ &= \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m} \end{aligned}$$

De donde la variable estandarizada

$$\frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{\sqrt{\sigma_x^2/n + \sigma_y^2/m}}$$

sigue una distribución normal estándar. Por consiguiente, si la hipótesis nula H_0: $\mu_x = \mu_y$ es cierta, el estadístico del contraste, TS (iniciales de *Test Statistic*), dado por

$$\text{TS} = \frac{\overline{X} - \overline{Y}}{\sqrt{\sigma_x^2/n + \sigma_y^2/m}} \tag{10.1}$$

seguirá una distribución normal estándar. Ahora bien, la probabilidad de que el valor absoluto de una variable aleatoria normal estándar Z sobrepase $z_{\alpha/2}$ es α; esto es,

$$P\{|Z| \geq z_{\alpha/2}\} = 2P\{Z \geq z_{\alpha/2}\} = \alpha$$

Así pues, ya que deseamos rechazar H_0 siempre que |TS| sea grande, se sigue que la conducta apropiada para contrastar

$$H_0: \mu_x = \mu_y \quad \text{frente} \quad H_1: \mu_x \neq \mu_y$$

a nivel α es

$$\begin{array}{ll} \text{Rechazar } H_0 & \text{si } |\text{TS}| \geq z_{\alpha/2} \\ \text{No rechazar } H_0 & \text{en otro caso} \end{array}$$

donde el estadístico del contraste, TS, está dado en la ecuación (10.1).

Una forma alternativa de llevar a cabo este contraste consiste en computar primero el valor del estadístico de contraste, TS; digamos que este valor es v. El p valor resultante para el contraste de H_0 frente a H_1 es la probabilidad de que el valor absoluto de una variable aleatoria normal estándar sea al menos tan grande como $|v|$. Así, si el valor de TS es v, entonces

$$p \text{ valor} = P\{|Z| \geq |v|\} = 2P\{Z \geq |v|\}$$

donde Z representa una variable aleatoria normal estándar.

Ejemplo 10.1 Se han propuesto dos nuevos métodos para producir un neumático. El fabricante cree que no existirán diferencias apreciables en los tiempos de vida de los neumáticos producidos por dichos métodos. Para contrastar la plausibilidad de esta hipótesis, se seleccionaron dos muestras: una de 9 neumáticos producidos con el método 1, y otra de 7 neumáticos producidos con el método 2. La primera muestra de neumáticos se prueba en carretera en un área A y la segunda en un área B. Por experiencia previa se sabe que el tiempo de vida de un neumático probado en cualquiera de dichas áreas es una variable aleatoria normal cuya media depende del neumático pero cuya varianza depende del área. Concretamente, se sabe que los tiempos de vida de los neumáticos probados en el área A son normales con una desviación típica igual a 3000 kilómetros, mientras que los probados en el área B tienen tiempos de vida que son normales con una desviación típica de 4000 kilómetros.

¿Los datos de la tabla 10.1 inducirán a que el productor rechace la hipótesis de que los tiempos de vida medios son los mismos para ambos tipos de neumáticos? Use un nivel de significación del 5%.

Solución Llamemos muestra X a los neumáticos probados en el área A, y muestra Y a aquellos probados en el área B. Para contrastar

$$H_0: \mu_x = \mu_y \quad \text{frente} \quad H_1: \mu_x \neq \mu_y$$

se necesita computar el valor del estadístico del contraste, TS. Ahora bien, las medias muestrales vienen dadas por

$$\overline{X} = 62{,}2444 \qquad \overline{Y} = 58{,}2714$$

Tabla 10.1 Vidas de los neumáticos en unidades de 1000 kilómetros

Neumáticos probados en A	Neumáticos probados en B	Neumáticos probados en A	Neumáticos probados en B
66,4	58,2	61,4	58,7
61,6	60,4	62,5	56,1
60,5	55,2	64,4	
59,1	62,0	60,7	
63,6	57,3		

Puesto que $n = 9$, que $m = 7$, $\sigma_x = 3$ y $\sigma_y = 4$, se ve que el valor del estadístico del contraste es

$$\text{TS} = \frac{62{,}2444 - 58{,}2714}{\sqrt{9/9 + 16/7}} = 2{,}192$$

Así pues, el p valor es igual a

$$p \text{ valor} = 2P\{Z \geq 2{,}192\} = 0{,}0284$$

y, por tanto, la hipótesis de medias iguales se rechaza a cualquier nivel de significación mayor o igual que 0,0284. En particular, se rechaza al nivel del 5% ($\alpha = 0{,}05$). ■

Si estuviéramos interesados en contrastar la hipótesis nula

$$H_0: \mu_x \leq \mu_y$$

frente a la alternativa unilateral

$$H_1: \mu_x > \mu_y$$

la hipótesis nula sólo se rechazará cuando el estadístico del contraste, TS, sea grande. Por tanto, en este caso el contraste al nivel de significación α consistirá en

Rechazar H_0	si $\text{TS} \geq z_\alpha$
No rechazar H_0	en otro caso

donde

$$\text{TS} = \frac{\bar{X} - \bar{Y}}{\sqrt{\sigma_x^2/n + \sigma_y^2/m}}$$

Equivalentemente, si el valor observado de TS es v, el p valor será

$$p \text{ valor} = P\{Z \geq v\}.$$

Ejemplo 10.2 Supongamos que el objetivo del experimento del ejemplo 10.1 sea intentar probar la hipótesis de que la vida media del primer conjunto de neumáticos sobrepasa la del segundo conjunto en más de 1000 kilómetros. ¿Los datos tienen la fuerza suficiente para corroborar esta hipótesis, por ejemplo, al nivel de significación del 5%?

Solución Denotemos por Y_i la vida del i-ésimo neumático del segundo conjunto, $i = 1,\ldots,7$. Si fijamos $W_i = Y_i + 1$, estaremos interesados en determinar si los datos nos permiten concluir que $\mu_x > \mu_w$, donde μ_x es la vida media de los neumáticos del primer conjunto y μ_w es la media de W_i. Para decidir esto, esta conclusión debería ser la hipótesis alternativa. Es decir, deberíamos contrastar

$$H_0: \mu_x \leq \mu_w \qquad \text{frente a} \qquad H_1: \mu_x > \mu_w$$

Dicho de otro modo, el rechazo de H_0 significaría una fuerte evidencia para validar la hipótesis de que la vida media del primer conjunto de neumáticos sobrepasa a la del segundo en más de 1000 kilómetros.

Para contrastar esta hipótesis calculamos el valor del estadístico del contraste, TS, teniendo cuidado en añadir 1 a los valores de los neumáticos probados en el área B que figuran en la tabla 10.1. Esto conduce a

$$\overline{X} = 62{,}2444 \qquad \overline{W} = 59{,}2714$$

y

$$\text{TS} = \frac{62{,}2444 - 59{,}2714}{\sqrt{9/9 + 16/7}} = 1{,}640$$

Puesto que se quiere rechazar H_0 cuando TS sea grande, el *p* valor es la probabilidad de que una normal estándar sobrepase 1,640. Esto es,

$$p \text{ valor} = P\{Z \geq 1{,}640\} = 0{,}0505$$

Así pues, aunque la evidencia está en gran medida a favor de la hipótesis alternativa, no es lo suficientemente fuerte para hacernos rechazar la hipótesis nula al nivel de significación del 5%. ■

La tabla 10.2 detalla tanto los contrastes bilaterales como los unilaterales presentados en esta sección.

Tabla 10.2 Contrastes de medias de dos poblaciones normales teniendo varianzas conocidas cuando las muestras son independientes

La media muestral, de una muestra de tamaño n procedente de una población normal con media μ_x y varianza conocida σ_x^2, es $\overline{X}$. La media muestral, de una muestra de tamaño m procedente de una segunda población normal con media μ_y y varianza conocida σ_y^2, es $\overline{Y}$. Las dos muestras son independientes.

H_0	H_1	Estadístico del contraste TS	Contraste a nivel de significación α	*p* valor si TS = *v*
$\mu_x = \mu_y$	$\mu_x \neq \mu_y$	$\dfrac{\overline{X} - \overline{Y}}{\sqrt{\sigma_x^2/n + \sigma_y^2/m}}$	Rechazar H_0 si $\lvert\text{TS}\rvert \geq z_{\alpha/2}$ No rechazar H_0 en otro caso	$2P\{Z \geq \lvert v\rvert\}$
$\mu_x \leq \mu_y$	$\mu_x > \mu_y$	$\dfrac{\overline{X} - \overline{Y}}{\sqrt{\sigma_x^2/n + \sigma_y^2/m}}$	Rechazar H_0 si $\text{TS} \geq z_\alpha$ No rechazar H_0 en otro caso	$P\{Z \geq v\}$

Problemas

1. Se lleva a cabo un experimento para contrastar la diferencia de efectividad de dos métodos de cultivo de trigo. En 12 pequeñas superficies de tierra se hace una siembra superficial y en otras 14 se realiza una siembra profunda. La producción media en las superficies de tierra del primer grupo es de 45,2 bushels, y en las superficies del segundo grupo es de 48,6 bushels. Suponga que se sabe que la siembra superficial hace que la producción en las superficies tenga una desviación típica de 0,8 bushels, mientras que la siembra profunda ocasiona que una desviación típica de 1,0 bushels.

   (a) ¿Son los datos consistentes, al nivel de significación del 5%, con la hipótesis de que la producción media es la misma con ambos métodos?

   (b) ¿Cuál es el p valor para este contraste de hipótesis?

2. Un método para medir el nivel de pH de una solución conduce a un valor de medida que se distribuye normalmente con media igual al valor real de pH y con desviación típica igual a 0,05. Un científico especializado en polución ambiental mantiene que dos soluciones diferentes provienen de la misma fuente. Si esto fuera así, los niveles de pH de las dos soluciones serían iguales. Para contrastar la plausibilidad de su tesis, se hacen 10 mediciones independientes del nivel de pH en ambas soluciones, con los siguientes datos resultantes:

Medidas de la solución A	Medidas de la solución B	Medidas de la solución A	Medidas de la solución B
6,24	6,27	6,26	6,31
6,31	6,25	6,24	6,28
6,28	6,33	6,29	6,29
6,30	6,27	6,22	6,34
6,25	6,24	6,28	6,27

   (a) ¿Desaprueban estos datos la tesis del científico? Use un nivel de significación del 5%.

   (b) ¿Cuál es el p valor?

3. Dos máquinas usadas para cortar acero están calibradas para cortar exactamente las mismas longitudes. Para contrastar esta hipótesis se utiliza cada máquina para cortar 10 piezas de acero. Posteriormente, se miden las piezas (con un error de medida despreciable). Suponga que los datos resultantes son los siguientes:

Máquina 1	Máquina 2	Máquina 1	Máquina 2
122,40	122,36	121,76	122,40
123,12	121,88	122,31	122,12
122,51	122,20	123,20	121,78
123,12	122,88	122,48	122,85

Supongamos que se sabe que la desviación típica de las longitudes de corte (hechas por cualquiera de las máquinas) es igual a 0,50.

(a) Contraste la hipótesis de que las máquinas están ajustadas al mismo valor, esto es, que las longitudes medias de sus cortes son iguales. Utilice un nivel de significación del 5%.

(b) Encuentre el p valor.

4. Los siguientes valores proceden de muestras independientes extraídas de dos poblaciones diferentes.

Muestra 1: 122, 114, 130, 165, 144, 133, 139, 142, 150
Muestra 2: 108, 125, 122, 140, 132, 120, 137, 128, 138

Sean μ_1 y μ_2 las respectivas medias de las dos poblaciones. Encuentre el p valor del contraste con hipótesis nula

$$H_0: \mu_1 \leq \mu_2$$

frente a la alternativa

$$H_1: \mu_1 > \mu_2$$

cuando las desviaciones típicas de las poblaciones son $\sigma_1 = 10$ y

(a) $\sigma_2 = 5$

(b) $\sigma_2 = 10$

(c) $\sigma_2 = 20$

5. En esta sección, se ha presentado el contraste de

$$H_0: \mu_x \leq \mu_y \qquad \text{frente a} \qquad H_1: \mu_x > \mu_y$$

Explique por qué no fue necesario presentar separadamente el contraste de

$$H_0: \mu_x \geq \mu_y \qquad \text{frente a} \qquad H_1: \mu_x < \mu_y$$

6. Un aparato usado por los astrónomos para medir distancias produce mediciones que tienen un valor medio igual a la distancia real al objeto observado y una desviación típica de 0,5 años luz. Un astrónomo está interesado en contrastar la hipótesis ampliamente aceptada de que un asteroide A está al menos tan cerca de la tierra como otro asteroide B. Para contrastar esta hipótesis, el astrónomo hizo 8 mediciones independientes del asteroide A, y 12 del B. Si la media de las medidas realizadas sobre el asteroide A fue de 22,4 años luz y la media de las realizadas sobre el asteroide B fue de 21,3, ¿será rechazada la hipótesis al nivel de significación del 5%? ¿Cuál es el p valor?

7. El valor recibido en una estación receptora de mensajes es igual al valor enviado más un error aleatorio que es normal, con media 0 y desviación típica 2. Se van a enviar dos mensajes, cada uno compuesto por un solo valor. Debido al error aleatorio, cada mensaje se enviará 9 veces. Previamente a la recepción, el receptor tiene la convicción de

que el valor del primer mensaje será menor o igual que el del segundo. ¿Se debería rechazar esta hipótesis si la media de los valores recibidos con respecto al mensaje 1 es 5,6 y la media de los recibidos referidos al mensaje 2 es 4,1? Utilice un nivel de significación del 1%.

8. Una gran compañía industrial lleva a cabo sus operaciones manufactureras a orillas de un gran río. Un oficial de la salud pública piensa que la compañía está incrementando el nivel de bifenil policlorinatado (BPC) del río mediante los vertidos de desechos tóxicos. Para obtener información, el oficial tomó 12 mediciones del agua de la parte del río próxima a la compañía y 14 mediciones en la orilla contraria. La media muestral de las 12 medidas tomadas cerca de la compañía fue de 32 partes por millón, y la media muestral del otro conjunto de 14 medidas fue de 22 partes por millón. Asuma que el valor de cada medida del agua es igual al valor real del nivel de BPC de la orilla del río donde se recogió la muestra más un error aleatorio debido al aparato de medición que es normal, con media 0 y desviación típica 8 partes por millón.

   (a) Con los datos dados y al nivel de significación del 5%, ¿podemos rechazar la hipótesis de que el nivel de BPC en la orilla de la compañía no es mayor que el nivel de de BPC en la orilla opuesta?

   (b) ¿Cuál es el p valor?

10.3 Contraste de igualdad de medias: varianzas desconocidas y tamaños muestrales grandes

En la sección anterior se ha supuesto que las varianzas poblacionales eran conocidas por el experimentador. Sin embargo, es mucho más habitual que estos parámetros sean desconocidos. Es decir, si la media de la población es desconocida, lo más probable es que la varianza también sea desconocida.

Supongamos de nuevo que tenemos dos muestras independientes $X_1,\ldots,X_n$ e $Y_1,\ldots,Y_m$ y que estamos interesados en contrastar una hipótesis relativa a sus medias μ_x y μ_y. Se asumirá en lo que sigue que las varianzas poblacionales σ_x^2 y σ_y^2 son desconocidas y, adicionalmente, que los tamaños muestrales n y m son grandes.

Para determinar el contraste apropiado en esta situación, se contará con que para tamaños de muestra grandes las varianzas muestrales igualarán aproximadamente las varianzas poblacionales. Así pues, parece razonable que podamos sustituir las varianzas muestrales S_x^2 y S_y^2 por las varianzas poblacionales y que usemos los análisis desarrollados en la sección anterior. Esto es, análogo al resultado de que

$$\frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{\sqrt{\sigma_x^2/n + \sigma_y^2/m}}$$

sigue una distribución normal estándar, parece previsible que para valores grandes de n y m, la variable aleatoria

$$\frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{\sqrt{S_x^2/n + S_y^2/m}}$$

seguirá una distribución aproximadamente normal estándar. Puesto que este resultado es realmente cierto, se sigue que podemos utilizar los mismos contrastes desarrollados en la sección 10.2 excepto que las varianzas muestrales se utilizan ahora en lugar de las varianzas poblacionales. Por ejemplo, el contraste a nivel de significación α de

$$H_0: \mu_x = \mu_y$$

frente a

$$H_1: \mu_x \neq \mu_y$$

nos hará rechazar H_0 cuando $|TS| \geq z_{\alpha/2}$, donde el estadístico del contraste, TS, ahora viene dado por

$$TS = \frac{\overline{X} - \overline{Y}}{\sqrt{S_x^2/n + S_y^2/m}}$$

Una forma equivalente de determinar el resultado del contraste consiste en obtener primero el valor del estadístico del contraste, TS, digamos que es v, para calcular después el p valor, dado por

$$p \text{ valor} = P\{|Z| \geq |v|\} = 2P\{Z \geq |v|\}$$

Igualmente, si se pretende contrastar la hipótesis unilateral

$$H_0: \mu_x \leq \mu_y$$

frente a

$$H_1: \mu_x > \mu_y$$

se utilizará el mismo estadístico que antes. El contraste será

Rechazar H_0	si $TS \geq z_\alpha$
Aceptar H_0	en otro caso

Equivalentemente, si el valor observado de TS es v, el p valor es

$$p \text{ valor} = P\{Z \geq v\}$$

Observaciones: Aún no se ha especificado lo grande que deberían ser n y m para que lo anterior sea válido. Una regla general a mano es que ambos tamaños muestrales sean cuanto menos 30, aunque valores de 20 o más son habitualmente suficientes.

Incluso cuando las distribuciones poblacionales subyacentes no sean normales, el teorema central del límite implica que las medias muestrales $\overline{X}$ e $\overline{Y}$ serán aproximadamente normales. Por esta razón, se pueden usar los anteriores contrastes entre medias poblacionales para distribuciones subyacentes arbitrarias siempre que los tamaños muestrales sean grandes. (De nuevo, unos tamaños muestrales de al menos 20 serán suficientes)

Ejemplo 10.3 Para contrastar la efectividad de una nueva medicación para rebajar el colesterol, se han dividido aleatoriamente a 100 voluntarios en dos grupos de 50 cada uno. A los miembros del primer grupo se les suministraron píldoras que contenían la medicación, mientras que a los miembros del segundo grupo, o grupo de *control*, se les suministraron píldoras que contenían lovastatin, uno de los medicamentos estándar para reducir el colesterol en la sangre. A todos los voluntarios se les indicó que tomaran las píldoras cada 12 horas durante los siguientes tres meses. Ninguno de los voluntarios sabía a qué grupo pertenecía.

Supongamos que el resultado de este experimento fue una reducción media de 8,2 con una varianza muestral de 5,4 en los niveles de colesterol en la sangre de aquéllos a los que se les suminitró el la medicación antigua, y una reducción media de 8,8 con una varianza muestral de 4,5 para aquellos que tomaron la medicación más moderna. ¿Prueban estos resultados, al nivel del 5%, que la nueva medicación es más efectiva que la antigua?

Solución Denotemos por μ_x la reducción media de colesterol de los voluntarios a quienes se les da la medicación nueva, y sea μ_y el valor equivalente de aquéllos a los que se les suministra el control. Si se quiere ver si los datos son suficientes para probar que $\mu_x > \mu_y$ se deberían utilizar para contrastar

$$H_0: \mu_x \leq \mu_y \qquad \text{frente a} \qquad H_1: \mu_x > \mu_y$$

El valor del estadístico del contraste es

$$TS = \frac{8{,}8 - 8{,}2}{\sqrt{4{,}5/50 + 5{,}4/50}} = 1{,}3484$$

Puesto que se trata de un contraste unilateral, en el que la hipótesis nula se rechazará cuando TS sea grande, el p valor es igual a la probabilidad de que una normal estándar (que sería la distribución aproximada de TS si $\mu_x = \mu_y$) sea al menos 1,3484. Esto es, el p valor con estos datos coincide con

$$p \text{ valor} = P\{Z \geq 1{,}3484\} = 0{,}089$$

Puesto que el p valor es mayor que 0,05, la evidencia no es lo suficientemente fuerte para establecer, al nivel de significación del 5%, que la nueva medicación es más efectiva que la antigua. ■

En el ejemplo 10.3, observe que se comparó la nueva medicación con la antigua, en lugar de compararla con un placebo. Ahora bien, si existe un tratamiento actualmente en uso, la medicación nueva se debería contrastar frente a éste. Esto es obvio en enfermedades muy serias para las que se prefiere concluir que la medicación nueva es mejor que la actual, en vez de concluir que la nueva es "mejor que nada".

Ejemplo 10.4 Un efecto bastante similar al efecto placebo se observa a menudo en experimentos industriales sobre el factor humano. Se ha observado que la productividad de un trabajador se incrementa habitualmente cuando ese trabajador es consciente de que está siendo monitorizado. Debido a que este fenómeno fue documentado y ampliamente publicitado tras algunos estudios sobre los incrementos de productividad llevados a cabo en la planta Howthorne de la compañía Western Electric, en ocasiones se le denomina *efecto Howthorne*. Para contrarrestar este efecto, los experimentos industriales hacen a menudo uso de un grupo de control.

Una consultora industrial ha sugerido una modificación del método existente para producir semiconductores. Ésta mantiene que esta modificación incrementará el número de semiconductores que puede producir un trabajador al día. Para contrastar la efectividad de sus ideas, el equipo de gestión ha planteado un pequeño estudio. Se ha dividido aleatoriamente a un conjunto de 50 trabajadores en dos grupos. A uno de ellos, formado por 30 trabajadores, se le entrenó para que afrontara la modificación propuesta por la consultora. El otro grupo, actuando como control, fue sometido a un entrenamiento con respecto a una modificación diferente. El equipo de gestión considera que ambas modificaciones son grosso modo iguales en complejidad de aprendizaje y en tiempo de implementación. Adicionalmente, el equipo de gestión está bastante seguro de que la modificación alternativa (a la propuesta por la consultora) no tendrá ningún efecto real sobre la productividad. A ninguno de los grupos se le indicó si estaba aprendiendo la propuesta de la consultora o no. Se monitorizó a los trabajadores durante un periodo de tiempo con los resultados siguientes.

Para los trabajadores que eran entrenados para la técnica de la consultora:

El número medio de semiconductores producidos por cada trabajador fue 242.

La varianza muestral fue 62,2.

Para los trabajadores del grupo de control:

El número medio de semiconductores producidos por cada trabajador fue 234.

La varianza muestral fue 58,4.

¿Estos datos son suficientes para probar que la modificación de la consultora incrementará la productividad?

Perspectiva histórica

La idea de usar una parte de una muestra como control es bastante antigüa. En el siglo XI, el médico árabe Avicena propuso unas reglas para la experimentación médica con sujetos humanos. Algunas de ellas implicaban el uso de controles. En 1626, Francis Bacon dio cuenta por escrito de los efectos de impregnar semillas de trigo con nueve mixturas diferentes, tales como el agua mezclada con excrementos de vaca, orina y diferentes tipos de vino, usando semillas no impregnadas como control. La mayor producción se consiguió con la impregnación de orina.

La primera exposición general sobre los experimentos que usaban controles fue realizada por Arthur Young. Él mantuvo que en los experimentos agrícolas siempre se tendría que comparar un tratamiento nuevo con otro conocido. En 1771 publicó sus ideas en el libro *Un curso sobre agricultura experimental.*

Solución Denotemos por μ_x el número medio de semiconductores que, durante el periodo de estudio, los trabajadores entrenados en el método de la consultora podrían producir. Igualmente, denotemos por μ_y el número medio producido por los trabajadores que utilizan la técnica alternativa. Para probar la idea de la consultora de que $\mu_x > \mu_y$ se necesitaría contrastar

$$H_0: \mu_x \leq \mu_y \qquad \text{frente a} \qquad H_1: \mu_x > \mu_y$$

Los datos son:

$$n = 30 \qquad m = 20$$
$$\bar{X} = 242 \qquad \bar{Y} = 234$$
$$S_x^2 = 62{,}2 \qquad S_y^2 = 58{,}4$$

Así pues, el valor del estadístico del contraste es

$$\text{TS} = \frac{242 - 234}{\sqrt{62{,}2/30 + 58{,}4/20}} = 3{,}58$$

De donde, el p valor de estos datos es

$$\text{valor } p = P\{Z \geq 3{,}58\} = 0{,}0002$$

En conclusión, los datos son suficientemente significativos para probar que la modificación de la consultora es más efectiva que la usada por el grupo de control. ■

Cuando se dispone de los datos individuales, en lugar de los estadísticos *sumariales*, las medias y las varianzas muestrales se pueden calcular manualmente o bien usando calculadoras o programas de ordenador similares al Programa 3-1. Estos valores se deberán usar después para determinar el valor del estadístico del contraste, TS. Finalmente, el p valor se puede obtener si se usa la tabla de probabilidad normal (tabla D.1 del Apéndice D).

Perspectiva histórica

El efecto Hawthorne ilustra el hecho de que la presencia de un observador puede afectar a la conducta de aquellos que están siendo observados. Como se ve en el ejemplo 10.4, el reconocimiento de este fenómeno no se consideró en la investigación realizada en la década de 1920 en la planta Hawthorne de Western Electric. Los investigadores intentaban determinar cómo se podía mejorar la productividad de los trabajadores de la planta. Sus estudios iniciales fueron diseñados para examinar los efectos que tenía la intensidad lumínica sobre la productividad de los trabajadores que ensamblaban componentes telefónicos. Se hicieron incrementos graduales en la luminosidad, y cada cambio condujo a productividades mayores. De hecho, la productividad continuó incrementándose incluso cuando la luminosidad se mantuvo anormalmente brillante. Más sorprendente aún fue el hecho de que la productividad continuó aumentando cuando la luminosidad se redujo.

Ejemplo 10.5 Contraste

$$H_0: \mu_x \leq \mu_y \qquad \text{frente a} \qquad H_1: \mu_x > \mu_y$$

con los datos siguientes:

X: 22, 21, 25, 29, 31, 18, 28, 33, 28, 26, 32, 35, 27, 29, 26
Y: 14, 17, 22, 18, 19, 21, 24, 33, 28, 22, 27, 18, 21, 19, 33, 31

Solución Un cálculo simple conduce a

$$\begin{aligned} n &= 15 & m &= 16 \\ \overline{X} &= 27{,}333 & \overline{Y} &= 22{,}938 \\ S_x^2 &= 21{,}238 & S_y^2 &= 34{,}329 \end{aligned}$$

De donde el valor de TS es

$$\text{TS} = \frac{4{,}395}{\sqrt{21{,}238/15 + 34{,}329/16}} = 2{,}33$$

Puesto que se trata de un contraste unilateral en el que la hipótesis nula se deberá rechazar sólo con valores grandes de TS, se tiene que

$$p \text{ valor} = P\{Z \geq 2{,}33\} = 0{,}01$$

Por consiguiente, la hipótesis de que la media de la población X no es mayor que la de la población Y se debería rechazar a niveles de significación mayores o iguales que 0,01. ■

La tabla 10.3 detalla los contrastes presentados en este apartado, tanto los bilaterales como los unilaterales.

Tabla 10.3 Contrastes de medias de dos poblaciones con varianzas desconocidas, cuando las muestras son independientes y los tamaños muestrales son grandes

La media muestral y la varianza muestral, de una muestra de tamaño n procedente de una población normal con una media μ_x y σ_x^2, son $\overline{X}$ y S_x^2, respectivamente. La media muestral y la varianza muestral, de una muestra de tamaño n procedente de una población normal con media μ_x y σ_y^2, son $\overline{Y}$ y S_y^2, respectivamente. Las dos muestras son independientes, y tanto n como m son como mínimo 20.

H_0	H_1	Estadístico del contraste TS	Contraste a nivel de significación α	p valor si TS = v
$\mu_x = \mu_y$	$\mu_x \neq \mu_y$	$\dfrac{\overline{X} - \overline{Y}}{\sqrt{S_x^2/n + S_y^2/m}}$	Rechazar H_0 si $\lvert TS \rvert \geq z_{\alpha/2}$ No rechazar H_0 en otro caso	$2P\{Z \geq \lvert v \rvert\}$
$\mu_x \leq \mu_y$	$\mu_x > \mu_y$	$\dfrac{\overline{X} - \overline{Y}}{\sqrt{S_x^2/n + S_y^2/m}}$	Rechazar H_0 si $TS \geq z_\alpha$ No rechazar H_0 en otro caso	$P\{Z \geq v\}$

Problemas

1. Un instituto está interesado en determinar si dos de sus profesores tienen la misma capacidad para preparar a estudiantes para un examen de geometría a escala estatal. A 70 estudiantes de geometría del semestre actual se les dividió aleatoriamente en dos grupos de 35. El profesor 1 enseñó geometría al primer grupo, y el profesor 2 al segundo. Al final del semestre, los estudiantes se sometieron al examen estatal con los siguientes resultados:

Clase del profesor 1	Clase del profesor 2
$\overline{X} = 72{,}6$	$\overline{Y} = 74{,}0$
$S_x^2 = 6{,}6$	$S_y^2 = 6{,}2$

   A partir de estos resultados, ¿se puede concluir que los profesores no tienen la misma capacidad para preparar a los estudiantes para el citado examen? Use un nivel de significación del 5%. Especifique las hipótesis nula y alternativa y el p valor resultante.

2. Los pesos muestrales (en libras) de los bebés nacidos en dos condados adyacentes al oeste de Pennsylvania conducen a los datos siguientes:

$$\begin{array}{ll} n = 53 & m = 44 \\ \overline{X} = 6{,}8 & \overline{Y} = 7{,}2 \\ S^2 = 5{,}2 & S^2 = 4{,}9 \end{array}$$

   Contraste la hipótesis de que los pesos medios de los recién nacidos son iguales en los dos condados. ¿Cuál es el p valor resultante? ¿Cómo explicaría sus conclusiones a una persona inteligente que aún no ha estudiado Estadística?

3. Un administrador de una gran estación termal tiene curiosidad por saber si, entre los clientes de la estación, las mujeres menores de 40 años visitan la estación con la misma frecuencia que las mujeres mayores de 40. Se eligieron dos muestras de 30 unidades cada una entre las clientas menores de 40 y entre las mayores de dicha edad, las cuales fueron monitorizadas durante el siguiente mes. El resultado fue que el grupo más joven tuvo una media muestral de 3,6 visitas con una desviación típica muestral de 1,3 visitas, mientras que la media muestral del otro grupo fue de 3,8 visitas con una desviación típica de 1,4. Use estos datos para contrastar la hipótesis de que el número medio de visitas de las población de clientas mayores de 40 es el mismo que el de la población de las clientas más jóvenes.

4. Usted está interesado en contrastar la hipótesis de que, por la mañana, el tiempo medio de viaje desde su casa al trabajo es igual al tiempo medio de viaje de vuelta del trabajo a casa por la tarde. Para ello, usted ha registrado los tiempos de 40 días laborables. Resultó que la media muestral de los viajes de ida al trabajo fue de 38 minutos, con una desviación típica

muestral de 4 minutos, y la media muestral de los viajes de vuelta a casa fue de 42 minutos, con una desviación típica muestral de 7 minutos.

(a) ¿Qué conclusión se puede sacar al nivel de significación del 5%?

(b) ¿Cuál es el *p* valor?

5. Se llevó a cabo el siguiente experimento para comparar las producciones de dos variedades de plantas de tomate. Tras una selección aleatoria, se plantaron en un terreno 36 plantas de cada variedad. Con la primera variedad se obtuvo una producción media muestral de 12,4 kilogramos por planta, con una desviación típica muestral de 1,6 kilogramos. Con la segunda variedad se obtuvo una producción media muestral de 14,2 kilogramos por planta, con una desviación típica muestral de 1,8 kilogramos. ¿Estos datos proporcionan suficiente evidencia para concluir que existen diferencias entre las producciones medias de las dos variedades? ¿A qué nivel de significación?

6. Se recogieron datos para determinar si existe una diferencia entre los resultados del test de inteligencia IQ de los estudiantes de las áreas rurales y los de las áreas urbanas del Estado de Michigan. Con una muestra aleatoria de 100 estudiantes urbanos se obtuvo una puntuación muestral media del test de 102,2 y una desviación típica de 11,8. Por su parte, una muestra de 60 estudiantes rurales produjo una puntuación media muestral de 105,3 con una desviación típica muestral de 10,6. ¿Los datos son suficientemente significativos, al nivel del 5%, para que rechacemos la hipótesis de que las puntuaciones medias de los estudiantes urbanos y rurales son iguales?

7. En el problema 6, ¿los datos son suficientemente significativos, al nivel del 1%, para concluir que la puntuación media de los estudiantes rurales de Michigan es superior a la de los estudiantes urbanos? ¿Cuáles son las hipótesis nula y alternativa?

8. En el problema 5 suponga que el investigador quisiera probar que la producción media de la segunda variedad es mayor que la de la primera. ¿Qué conclusiones se hubieran sacado? Use un nivel de significación del 5%.

9. Una compañía debe decidir entre dos suministradores de bombillas. La dirección ha decidido hacer el pedido al suministrador A, a menos que "se pueda probar" que la vida media de las bombillas del suministrador B es superior. Con un chequeo de 28 bombillas de A y de 32 bombillas de B se obtuvieron los siguientes datos del número de horas de uso que soportó cada bombilla:

A: 121, 76, 88, 103, 96, 89, 100, 112, 105, 101, 92, 98, 87, 75, 111, 118, 121, 96, 93, 82, 105, 78, 84, 96, 103, 119, 85, 84
B: 127, 133, 87, 91, 81, 122, 115, 107, 109, 89, 82, 90, 81, 104, 109, 110, 106, 85, 93, 90, 100, 122, 117, 109, 98, 94, 103, 107, 101, 99, 112, 90

Al nivel de significación del 5%, ¿a qué suministrador se debería elegir? Especifique las hipótesis que se deben contrastar y el *p* valor resultante.

10. Un administrador de una escuela de negocios mantiene que el salario medio de sus graduados, después de 10 años, es al menos 5000 dólares mayor que el de los graduados similares de una institución rival. Para estudiar esta hipótesis, se seleccionó una muestra de 50 estu-

diantes que se habían graduado 10 años antes, y sus salarios fueron registrados. También se extrajo una muestra similar de estudiantes de la institución rival. Supongamos que se obtuvieron los resultados siguientes:

Escuela de negocios	Institución rival
$n = 50$	$m = 50$
$\overline{X} = 85{,}2$	$\overline{Y} = 74{,}8$
$S_x^2 = 26{,}4$	$S_y^2 = 24{,}5$

(a) Para determinar si los datos corroboran la idea del administrador, ¿cuáles deberían ser las hipótesis nula y alternativa?

(b) ¿Cúal es el p valor resultante?

(c) ¿Qué conclusiones se pueden extraer?

11. Recientemente se ha intentado verificar si las mujeres de una cierta industria están siendo discriminadas negativamente, en lo concerniente a los salarios. Para estudiar este hecho, un investigador aceptado por el juzgado extrajo una muestra aleatoria de empleados con 8 o más años de experiencia y con un historial de empleo estable durante este tiempo. Con una unidad de salarios de 1 dólar, resultaron los siguientes salarios por hora:

Trabajadores mujeres	Trabajadores hombres
Tamaño muestral: 55	Tamaño muestral: 72
Media muestral: 10,80	Media muestral: 12,20
Varianza muestral: 0,90	Varianza muestral: 1,1

(a) ¿Qué hipótesis se deberían contrastar?

(b) ¿Cuál es el p valor resultante?

(c) ¿Qué prueba esto?

12. El siguiente extracto de datos se obtuvo de la comparación de los contenidos de plomo en pelos humanos recogidos de individuos adultos que murieron entre 1880 y 1920 y de adultos actuales. Los datos se registraron en unidades de microgramos, es decir de millonésimas de gramo.

	1880–1920	Hoy
Tamaño muestral	30	100
Media muestral	48,5	26,6
Desviación típica muestral	14,5	12,3

(a) ¿Establecen estos datos, al nivel de significación del 1%, que el contenido de plomo en pelos humanos es menor hoy que entre los años 1880 y 1920? Formule claramente cuáles son las hipótesis nula y alternativa.

(b) ¿Cuál es el p valor para las hipótesis contrastadas en la parte (a)?

13. Se dividió aleatoriamente a un grupo de 40 trabajadores en dos conjuntos de 20 cada uno. Cada conjunto empleó 2 semanas en un programa de autoaprendizaje que fue diseñado para aprender una nueva técnica de producción. Un supervisor acompañó al primer conjunto de trabajadores y su único trabajo fue chequear que todos los trabajadores prestaran atención. El segundo grupo realizó el curso sin supervisor. Cuando finalizó el programa se examinó a los trabajadores. Los resultados fueron los siguientes:

	Media muestral	**Desviación típica muestral**
Grupo supervisado	70,6	8,4
Grupo no supervisado	77,4	7,4

(a) Contraste la hipótesis nula de que la supervisión no tiene efecto sobre la actuación de los trabajadores. Use un nivel de significación del 1%.

(b) ¿Cuál es el p valor resultante?

(c) ¿Qué conclusión se puede obtener sobre el resultado de la supervisión?

10.4 Contraste de igualdad de medias: contrastes con muestras pequeñas cuando las varianzas poblacionales son desconocidas e iguales

Supongamos que se dispone de muestras independientes procedentes de dos poblaciones normales:

$$X_1, \ldots, X_n \qquad \text{y} \qquad Y_1, \ldots, Y_m$$

y que se está interesado en contrastar hipótesis relativas a sus respectivas medias muestrales μ_x y μ_y. A diferencia de las secciones anteriores, no se supondrá ni que las varianzas poblacionales son conocidas ni que los tamaños muestrales son necesariamente grandes.

En muchas situaciones es razonable suponer que las varianzas poblacionales σ_x^2 y σ_y^2 son aproximadamente iguales, incluso aunque sean desconocidas. Así pues, asumamos que son iguales y denotemos su valor común como σ^2. Esto es, supongamos que

$$\sigma_x^2 = \sigma_y^2 = \sigma^2$$

Para contrastar la hipótesis nula

$$\text{H}_0\text{: } \mu_x = \mu_y \qquad \text{frente a} \qquad \text{H}_1\text{: } \mu_x \neq \mu_y$$

cuando las varianzas poblacionales son iguales partimos del hecho, demostrado en la sección 10.2, de que

$$\frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{\sqrt{\sigma_x^2/n + \sigma_y^2/m}}$$

sigue una distribución normal estándar.

Así pues, puesto que $\sigma_x^2 = \sigma_y^2 = \sigma^2$, se observa que, cuando H_0 es cierta (y por tanto $\mu_x - \mu_y = 0$), $(\overline{X} - \overline{Y}) / \sqrt{\sigma^2/n + \sigma^2/m}$ sigue una distribución normal estándar. Esto es,

Si H_0 es cierta

$$\frac{\overline{X} - \overline{Y}}{\sqrt{\sigma^2/n + \sigma^2/m}} \tag{10.2}$$

sigue una distribución normal estándar.

El resultado precedente no se puede emplear directamente para contrastar la hipótesis nula de medias iguales puesto que involucra el parámetro desconocido σ^2. Como resultado, primero se obtendrá un estimador de σ^2 y después se determinará el efecto que tiene reemplazar σ^2 por su estimador sobre la distribución del estadístico (10.2).

Para obtener un estimador de σ^2 se hará uso del hecho de que las varianzas muestrales S_x^2 y S_y^2 estiman ambas la varianza poblacional común σ^2. Resulta, pues, natural combinar, *o poolear* (del término inglés *pool*), estos dos estimadores. En otras palabras, es natural considerar una media ponderada de las dos varianzas muestrales. Para determinar los pesos adecuados que se deben asignar a cada varianza muestral, recordemos que la varianza muestral de una muestra de tamaño, digamos, k tiene $k-1$ grados de libertad asociados a ella. De aquí se ve que S_x^2 tiene $n-1$ grados de libertad asociados, mientras que S_y^2 tiene $m-1$. Así pues, usaremos un estimador combinado que pondera S_x^2 con un peso de $(n-1)/(n-1+m-1)$ y S_y^2 con un peso de $(m-1)/(n-1+m-1)$.

Definición

El estimador S_p^2 definido por

$$S_p^2 = \frac{n-1}{n+m-2} S_x^2 + \frac{m-1}{n+m-2} S_y^2$$

Se llama *estimador combinado* (o "*pooleado*") de σ^2.

Observe que cuanto mayor es el tamaño muestral, mayor es el peso dado a su varianza muestral en la estimación de σ^2. Además, el estimador combinado tendrá asociados $n - 1 + m - 1 = n + m - 2$ grados de libertad.

Si en la expresión (10.2) se reemplaza σ^2 por su estimador combinado S_p^2, se puede demostrar que, cuando H_0 es cierta, el estimador resultante sigue una distribución t con $n + m - 2$ grados de libertad. [Esto es directamente análogo a lo que ocurre con la varianza muestral S: a saber, esta sustitución transforma la variable aleatoria $\sqrt{n}(\overline{X} - \mu)/\sigma$ que se distribuye como una normal estándar, en la variable aleatoria $\sqrt{n}(\overline{X} - \mu)/S$, que sigue ahora una t con $n - 1$ grados de libertad.]

De lo anterior se ve que para contrastar

$$H_0: \mu_x = \mu_y \qquad \text{frente a} \qquad H_1: \mu_x \neq \mu_y$$

se debería computar primero el valor del estadístico

$$TS = \frac{\overline{X} - \overline{Y}}{\sqrt{S_p^2(1/n + 1/m)}}$$

El contraste al nivel de significación α actuará, pues, como sigue

$$\begin{array}{ll} \text{Rechazar } H_0 & \text{si } |TS| \geq t_{n+m-2,\,\alpha/2} \\ \text{No rechazar } H_0 & \text{en otro caso} \end{array}$$

Alternativamente, se puede ejecutar el contraste determinando el p valor. Si el valor de TS observado resulta ser v, el p valor resultante del contraste de H_0 frente a H_1 viene dado por

$$\begin{aligned} p \text{ valor} &= P\{|T_{n+m-2}| \geq |v|\} \\ &= 2P\{T_{n+m-2} \geq |v|\} \end{aligned}$$

donde T_{n+m-2} es una variable aleatoria t con $n + m - 2$ grados de libertad.

Si se está interesado en contrastar las hipótesis unilaterales

$$H_0: \mu_x \leq \mu_y \qquad \text{frente a} \qquad H_1: \mu_x > \mu_y$$

se debería rechazar H_0 cuando se produzcan valores grandes de TS. Así pues, el contraste a nivel de significación α actuará como sigue:

$$\begin{array}{ll} \text{Rechazar } H_0 & \text{si } |TS| \geq t_{n+m-2,\,\alpha} \\ \text{No rechazar } H_0 & \text{en otro caso} \end{array}$$

Si el valor del estadístico del contraste, TS, es v, el p valor viene dado por

$$p \text{ valor} = P\{T_{n+m-2} \geq v\}$$

El Programa 10-1 computa el valor del estadístico del contraste y el correspondiente *p* valor tanto para los contrastes unilaterales como para los bilaterales.

Ejemplo 10.6 Veintidós voluntarios de un instituto de investigación del resfriado adquirieron la enfermedad tras ser expuestos a varios virus. Se seleccionó aleatoriamente a 10 de ellos y, cuatro veces al día, se les suministraron píldoras que contenían 1 gramo de vitamina C. Al grupo de control, compuesto por los 12 voluntarios restantes, se les suministró un placebo que tenía una apariencia y un sabor similares a las píldoras de vitamina C. Un doctor, que desconocía qué voluntarios habían recibido la vitamina C o el placebo, dictaminó posteriormente el momento en el que cada paciente dejó de padecer el resfriado, y se registraron las duraciones de la enfermedad en cada caso.

Al final del experimento, se obtuvieron los datos siguientes:

Tratados con vitamina C	Tratados con placebo	Tratados con vitamina C	Tratados con placebo
5,5	6,5	7,5	7,5
6,0	6,0	5,5	6,5
7,0	8,5	7,0	7,5
6,0	7,0	6,5	6,0
7,5	6,5		8,5
6,0	8,0		7,0

¿Prueban estos datos que tomar 4 gramos de vitamina C diariamente reduce el tiempo de duración del resfriado? ¿A qué nivel de significación?

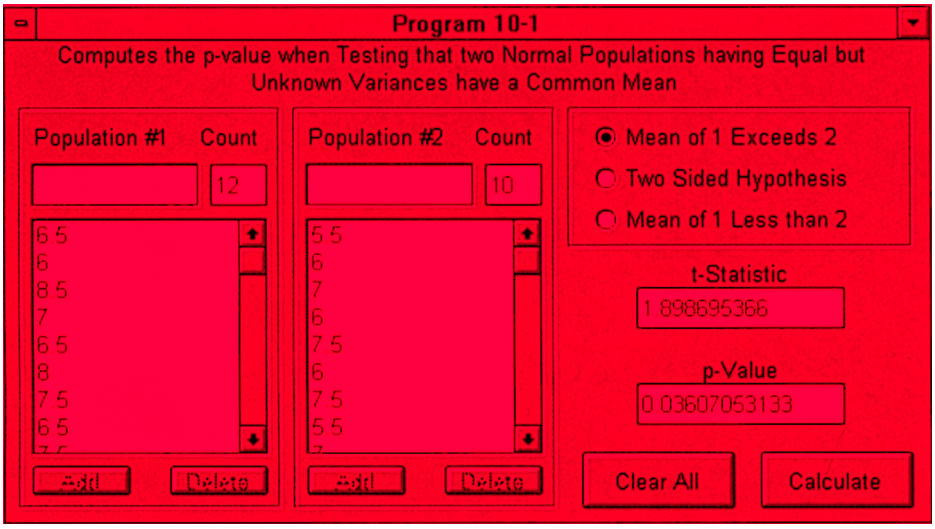
Solución Para probar la anterior hipótesis, se necesitaría rechazar la hipótesis nula al contrastar

$$H_0: \mu_p \leq \mu_c \qquad \text{frente a} \qquad H_1: \mu_p > \mu_c$$

donde μ_c es el tiempo medio que dura un resfriado cuando se toman las tabletas de vitamina C y μ_p es el equivalente de tiempo medio cuando se toma el placebo. Si suponemos que la varianza de las duraciones del resfriado es la misma para los pacientes tratados con vitamina C que para los tratados con placebo, hemos contrastado las hipótesis con el programa 10-1. Este programa computa el *p* valor cuando se contrasta si dos poblaciones normales con varianzas iguales y desconocidas tienen medias iguales.

Los valores de la muestra 1 son los siguientes: 6,5, 6, 8,5, 7, 6,5, 8, 7,5, 6,5, 7,5, 6, 8,5 y 7.

Los valores de la muestra 2 son los siguientes: 5,5, 6, 7, 6, 7,5, 6, 7,5, 5,5, 7 y 6,5.



El Programa 10-1 computa el valor del estadístico t como 1,898695366.

Cuando se introducen los valores de los datos en el Programa 10-1, asegúrese de que la hipótesis alternativa no es bilateral, sino que, por el contrario, es que la media de la muestra 1 sobrepasa a la de la muestra 2.

Consiguientemente, el programa computa el p valor como 0,03607053133.

En consecuencia, H_0 se debería rechazar al nivel de significación del 5%.

Naturalmente, si no se quisiera ejecutar el Programa 10-1, se podría llevar a cabo el contraste calculando primero los valores de los estadísticos $\overline{X}$, $\overline{Y}$, S_x^2, S_Y^2 y S_p^2, donde la muestra X se corresponde con los pacientes que reciben el placebo, y la muestra Y con los que reciben la vitamina C. Con estos cálculos se obtienen los valores:

$$\overline{X} = 7{,}125 \qquad \overline{Y} = 6{,}450$$
$$S_x^2 = 0{,}778 \qquad S_y^2 = 0{,}581$$

Por consiguiente,

$$S_p^2 = \frac{11}{20} S_x^2 + \frac{9}{20} S_y^2 = 0{,}689$$

y el valor del estadístico del contraste es

$$\text{TS} = \frac{0{,}675}{\sqrt{0{,}689(1/12 + 1/10)}} = 1{,}90$$

Puesto que, de la tabla D.2, $t_{20,0,05} = 1{,}725$, se rechaza la hipótesis nula al nivel de significación del 5%. Esto es, existe una evidencia significativa, al nivel del 5%, para afirmar que la vitamina C reduce el tiempo medio de duración del resfriado. ■

Tabla 10.4 Contrastes de medias de dos poblaciones con varianzas desconocidas, aunque iguales cuando las muestras son independientes

La media muestral y la varianza muestral, de una muestra de tamaño n procedente de una población normal con media μ_x y σ^2, son $\overline{X}$ y S_x^2. Y la media muestral y la varianza muestral, de una muestra de tamaño n procedente de una segunda población normal con media μ_x y σ^2, son $\overline{Y}$ y S_y^2. Las dos muestras son independientes.

$$S_p^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}$$

H_0	H_1	Estadístico del contraste TS	Contraste a nivel de significación α	p valor si TS = v
$\mu_x = \mu_y$	$\mu_x \neq \mu_y$	$\frac{\overline{X} - \overline{Y}}{\sqrt{S_p^2(1/n + 1/m)}}$	Rechazar H_0 si $\lvert TS\rvert \geq t_{n+m-2,\,\alpha/2}$ No rechazar H_0 en otro caso	$2P\{T_{n+m-2} \geq \lvert v\rvert\}$
$\mu_x \leq \mu_y$	$\mu_x > \mu_y$	$\frac{\overline{X} - \overline{Y}}{\sqrt{S_p^2(1/n + 1/m)}}$	Rechazar H_0 si $TS \geq t_{n+m-2,\alpha}$ No rechazar H_0 en otro caso	$P\{T_{n+m-2} \geq v\}$

La tabla 10.4 detalla tanto los contrastes bilaterales como los unilaterales presentados en esta sección.

Problemas

En los siguientes problemas se asume que las distribuciones poblacionales son normales y tienen varianzas iguales.

1. Se seleccionó aleatoriamente a veinticinco varones con edades comprendidas entre 25 y 30 años, para participar en un muy conocido estudio del corazón llevado a cabo en Flamingham, Massachussets. Entre ellos, 11 eran fumadores, y 14 no. Los siguientes datos indican las lecturas hechas de sus presiones sanguíneas sistólicas:

Fumadores	No fumadores	Fumadores	No fumadores
124	130	131	127
134	122	133	135
136	128	125	120
125	129	118	122
133	118		120
127	122		115
135	116		123

¿Indican estos datos que existen, al nivel de significación del 1%, diferencias entre las medias de la presiones sanguíneas sistólicas de las poblaciones representadas por ambos grupos? Si no existen estas diferencias, ¿qué se puede decir al nivel de significación del 5%?

2. Se ha diseñado un estudio para conocer cómo cambian las dietas de las mujeres durante el verano y el invierno. Se observó a un grupo aleatorio de 12 mujeres durante el mes de julio para determinar qué porcentaje de calorías en sus dietas provenían de grasas. Observaciones similares se hicieron en el mes de enero sobre un grupo distinto de 12 mujeres seleccionadas aleatoriamente. Suponga que los resultados fueron los siguientes:

   Julio: 32,2, 27,4, 28,6, 32,4, 40,5, 26,2, 29,4, 25,8, 36,6, 30,3, 28,5, 32,0
   Enero: 30,5, 28,4, 40,2, 37,6, 36,5, 38,8, 34,7, 29,5, 29,7, 37,2, 41,5, 37,0

   Contraste la hipótesis de que la media de los porcentajes de consumo de grasas son iguales en ambos meses. Use los niveles de significación del:

   (a) 5%

   (b) 1%

3. Una organización de consumidores ha comparado el tiempo que tarda en elimianr el dolor un medicamento genérico con el tiempo que tarda otro medicamento de marca. Con cada uno de ellos se chequearon nueve casos, con los siguientes datos resultantes:

   Genérico: 14,2, 14,7, 13,9, 15,3, 14,8, 13,6, 14,6, 14,9, 14,2
   De marca: 14,3, 14,9, 14,4, 13,8, 15,0, 15,1, 14,4, 14,7, 14,9

   (a) ¿Establecen estos datos que el medicamento de marca elimina el dolor mas rápidamente al nivel de significación del 5%?

   (b) ¿Y al nivel de significación del 10%?

4. Para averiguar los hábitos de alimentación de los murciélagos se marcó y se siguió por radar a un grupo de 22 murciélagos. De éstos, 12 eran hembras y 10 eran machos. Se anotaron las distancias (en metros) recorridas entre dos periodos de alimentación para cada uno de los murciélagos, y se obtuvieron los siguientes estadísticos sumariales:

Murciélagos hembras	**Murciélagos machos**
$n = 12$	$m = 10$
$\overline{X} = 180$	$\overline{Y} = 136$
$S_x = 92$	$S_y = 86$

   Contraste la hipótesis de que las distancias recorridas son las mismas para las poblaciones de murciélagos machos que para las hembras. Use un nivel de significación del 5%.

5. Para determinar la efectividad de un nuevo método de enseñanza de lectura para niños pequeños, se dividió aleatoriamente a 20 niños que no sabían leer en dos grupos de 10 niños cada uno. Al primer grupo se le enseñó a leer por un método estándar y al segundo grupo

por el método experimental. Al final del cuatrimestre, los alumnos fueron sometidos a un examen de lectura, resultaron los siguientes estadísticos sumariales:

Estudiantes con el método estándar	**Estudiantes con el método experimental**
Puntuación media = 65,6	Puntuación media = 70,4
Desviación típica = 5,4	Desviación típica = 4,8

¿Tienen estos datos la fuerza suficiente para probar, al nivel de significación del 5%, que las puntuaciones con el método experimental son más altas?

6. Vuelva a hacer el problema 2 de la sección 10.3, suponiendo que las varianzas poblacionales son iguales.

   (a) ¿Se rechazaría la hipótesis nula al nivel de significación del 5%?

   (b) ¿Cuál es el p valor comparado con el obtenido previamente?

7. Para analizar cómo afecta la dieta sobre el padecimiento de una enfermedad diverticular, se sometieron a estudio 20 vegetarianos, 6 de los cuales tenían la enfermedad. Se determinó el total de fibra consumida diariamente por cada uno de los individuos, con los resultados siguientes:

Con la enfermedad	**Sin la enfermedad**
$n = 6$	$m = 14$
$\overline{X} = 26{,}8$ gramos	$\overline{Y} = 42{,}5$ gramos
$S_x = 9{,}2$ gramos	$S_y = 9{,}5$ gramos

Contraste la hipótesis de que el consumo de fibra diaria es igual para las poblaciones de vegetarianos que sufren la enfermedad diverticular que para los que no la padecen. Use un nivel de significación del 5%.

8. Es bien conocido que los habitantes del extrarradio de Los Ángeles conducen diariamente más millas que los habitantes del extrarradio de la bahía de San Francisco. Para comprobar si este "hecho" es realmente cierto se extrajeron dos muestras aleatorias de 20 habitantes, respectivamente, de los extrarradios de Los Ángeles y de la bahía de San Francisco, y sus hábitos de conducción fueron, después, monitorizados. Resultaron los siguientes datos respecto al número medio y a la desviación típica de las millas conducidas.

Extrarradio de Los Ángeles	**Extrarradio de San Francisco**
$\overline{X} = 57{,}4$	$\overline{X} = 52{,}8$
$S_x = 12.4$	$S_y = 13.8$

¿Prueban estos datos la hipótesis de que la distancia media de conducción de los habitantes del extrarradio de Los Ángeles sobrepasa a la de los habitantes del extrarradio de San Francisco? Use un nivel de significación del:

(a) 10%

(b) 5%

(c) 1%

9. Los siguientes datos son los resultados de dos muestras independientes procedentes de dos poblaciones diferentes:

X: 10,3, 10,4, 11,3, 13,5, 12,7, 11,1, 10,9, 9,7, 14,5, 13,3
Y: 12,4, 11,7, 13,5, 12,9, 13,4, 15,5, 16,3, 13,7, 14,3

Contraste la hipótesis nula de que las medias de las dos poblaciones son iguales frente a la alternativa de que son desiguales, al nivel de significación del:

(a) 10%

(b) 5%

(c) 1%

10. Un gerente está considerando institucionalizar un descanso adicional de 15 minutos si se puede probar que ello reduce el número de errores que cometen los empleados. El gerente divide una muestra de 20 empleados en dos grupos de 10 cada uno. Los miembros del primer grupo continúan con el mismo horario de trabajo habitual; a los miembros del otro grupo se les otorga un descanso de 15 minutos adicional. Los siguientes datos reflejan el número total de errores cometidos por cada uno de los trabajadores en los siguientes 20 días de trabajo.

Grupo con descanso: 8, 7, 5, 8, 10, 9, 7, 8, 4, 5
Grupo sin descanso: 7, 6, 14, 12, 13, 8, 9, 6, 10, 9

Contraste la hipótesis de que el descanso no reduce el número medio de errores al nivel de significación del 5%. ¿Cuál es la conclusión?

10.5 Contrastes de la *t* con muestras apareadas

Supongamos que $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_n$ son muestras del mismo tamaño procedentes de dos poblaciones normales diferentes con medias μ_x y μ_y, respectivamente. En ciertas situaciones existirá una relación entre los valores de los datos X_i e Y_i. Debido a esta relación, los pares de valores X_i, Y_i, $i = 1, \ldots, n$ no serán independientes; en consecuencia, no será posible utilizar los métodos de las secciones precedentes para contrastar hipótesis relativas a μ_x y μ_y.

Ejemplo 10.7 Suponga que deseamos averiguar el efecto que tiene un aditivo de la gasolina que se ha desarrollado recientemente sobre la distancia recorrida por unidad de

carburante. Para obtener información, han sido seleccionados siete coches, y las millas (por galón de gasolina) se registraron posteriormente. Para cada coche, esto se hizo usando gasolina sin aditivo y con aditivo. Los resultados se muestran a continuación:

Coche	Kilometraje sin aditivo	Kilometraje con aditivo
1	24,2	23,5
2	30,4	29,6
3	32,7	32,3
4	19,8	17,6
5	25,0	25,3
6	24,9	25,4
7	22,2	20,6

Por ejemplo, el coche 1 consiguió recorrer 24,2 millas por galón usando gasolina sin aditivo, y solamente 23,5 millas por galón usando gasolina con aditivo; mientras que el coche 4 consiguió recorrer 19,8 millas por galón usando gasolina sin aditivo, y 17,6 millas por galón usando gasolina con aditivo.

Ahora bien, es fácil ver que dos factores determinan la distancia por galón recorrida por los coches. El primero es si la gasolina incluye o no el aditivo, y el segundo es el propio coche. Por este motivo, no se pueden tratar las dos muestras como si fueran independientes, sino que se deberían considerar como datos apareados. ■

Supongamos que se quiere contrastar

$$H_0: \mu_x = \mu_y \qquad \text{frente a} \qquad H_1: \mu_x \neq \mu_y$$

donde las dos muestras consisten en datos apareados X_i, Y_i, $= 1, \ldots, n$. Se puede contrastar la hipótesis de que las dos medias poblacionales son iguales si se observan las diferencias entre los datos de manera apareada. Esto es, hagamos

$$D_i = X_i - Y_i \qquad i = 1, \ldots, n$$

Ahora bien,

$$E[D_i] = E[X_i] - E[Y_i]$$

también, si $\mu_d = E[D_i]$,

$$\mu_d = \mu_x - \mu_y$$

La hipótesis nula de que $\mu_x = \mu_y$ es, por tanto, equivalente a la hipótesis de que $\mu_d = 0$. En consecuencia, se puede contrastar la hipótesis de que las medias poblacionales son iguales si se contrasta

$$H_0: \mu_d = 0 \qquad \text{frente a} \qquad H_1: \mu_d \neq 0$$

Si asumimos que las variables aleatorias $D_1, \ldots, D_n$ constituyen una muestra procedente de una población normal, se puede contrastar la anterior hipótesis nula mediante el contraste de la t descrito en la sección 9.4. Esto es, si $\overline{D}$ y S_d denotan la media muestral y la desviación típica muestral, respectivamente, el estadístico del contraste, TS, viene dado por

$$\text{TS} = \sqrt{n}\,\frac{\overline{D}}{S_d}$$

El contraste a nivel de significación α consistirá en

Rechazar H_0	si $\lvert\text{TS}\rvert \geq t_{n-1,\,\alpha/2}$
No rechazar H_0	en otro caso

donde el valor $t_{n-1\ \alpha/2}$ se puede obtener de la tabla D.2.

Equivalentemente, se puede realizar el contraste computando el valor, digamos ν, del estadístico del contraste, TS, y después obteniendo el p valor dado por

$$p \text{ valor} = P\{|T_{n-1}| \geq |\nu|\} = 2P\{T_{n-1} \geq |\nu|\}$$

donde T_{n-1} es una variable aleatoria t con $n - 1$ grados de libertad. Si se dispone de un ordenador personal, se puede usar el Programa 9-1 para determinar el valor del estadístico del contraste y el p valor resultante. Los valores que se deben introducir en este programa son $D_1, D_2, \ldots, D_n$ y el valor de μ_0 (el valor de la hipótesis nula para la media de D) que, en este caso, debe ser 0.

Ejemplo 10.8 Con los datos del ejemplo 10.7, contraste, al nivel de significación del 5%, la hipótesis nula de que el aditivo no modifica el número medio de millas conducidas por galón de gasolina.

Solución Si no se cree conveniente ejecutar el Programa 9-1, se pueden calcular primero las diferencias D_i, y luego los estadísticos sumariales $\overline{D}$ y S_d. Las diferencias citadas son

$$0{,}7,\ 0{,}8,\ 0{,}4,\ 2{,}2,\ -0{,}3,\ -0{,}5,\ 1{,}6$$

de donde resultan los estadísticos

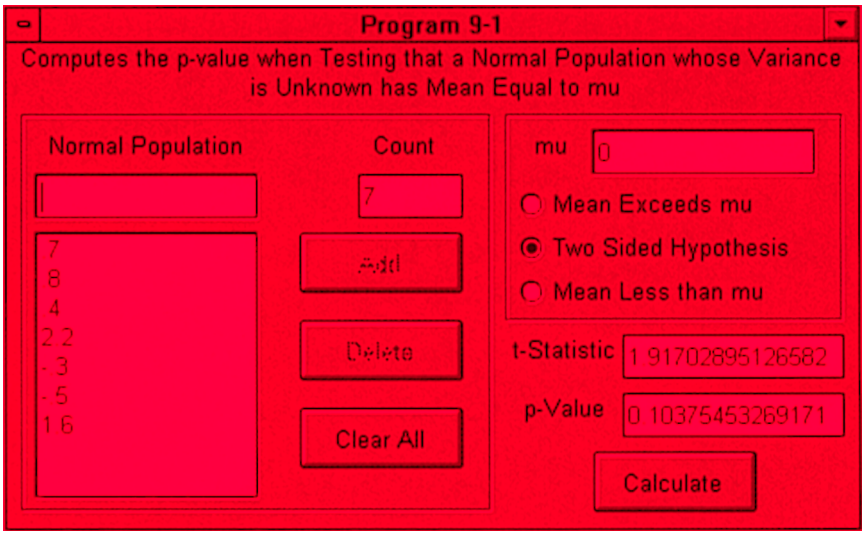
$$\overline{D} = 0{,}7 \qquad S_d = 0{,}966$$

Por consiguiente, el valor del estadístico del contraste resultante es

$$\text{TS} = \frac{\sqrt{7}(0{,}7)}{0{,}966} = 1{,}917$$

Puesto que, de la tabla D.2, $t_{6,\,0{,}025} = 2{,}447$, al nivel del 5%, no se rechaza la hipótesis de que la distancia media recorrida por galón es idéntica tanto si la gasolina usada contiene el aditivo como si no.

Si se dispone de un ordenador personal, se puede resolver el problema mediante el Programa 9-1. Éste conduce a lo siguiente:



Por tanto, la hipótesis nula no sería ni siquiera rechazada al nivel de significación del 10%. ■

Los contrastes unilaterales relativos a las dos medias poblacionales se obtienen de la misma forma. Por ejemplo, para contrastar

$$H_0: \mu_x \leq \mu_y \qquad \text{frente a} \qquad H_1: \mu_x > \mu_y$$

se pueden utilizar los datos $D_1, \ldots, D_n$ y contrastar después las hipótesis

$$H_0: \mu_d \leq 0 \qquad \text{frente a} \qquad H_1: \mu_d > 0$$

De nuevo con el estadístico del contraste

$$\text{TS} = \sqrt{n}\,\frac{\overline{D}}{S_d}$$

el contraste a nivel de significación α consistirá en

$$\begin{array}{ll} \text{Rechazar } H_0 & \text{si TS} > t_{n-1,\,\alpha} \\ \text{No rechazar } H_0 & \text{en otro caso} \end{array}$$

Equivalentemente, si el valor de TS es v, el p valor será

$$p \text{ valor} = P\{T_{n-1} \geq v\}$$

Se puede utilizar de nuevo el Programa 9-1 para determinar el valor del estadístico del contraste y el p valor resultante. (Si se conocen los estadísticos sumariales $\overline{D}$ y S_d el p valor se

puede obtener calculando ν, el valor del estadístico del contraste, y después ejecutando el Programa 8-1 para determinar $P\{T_{n-1} \geq \nu\}$.)

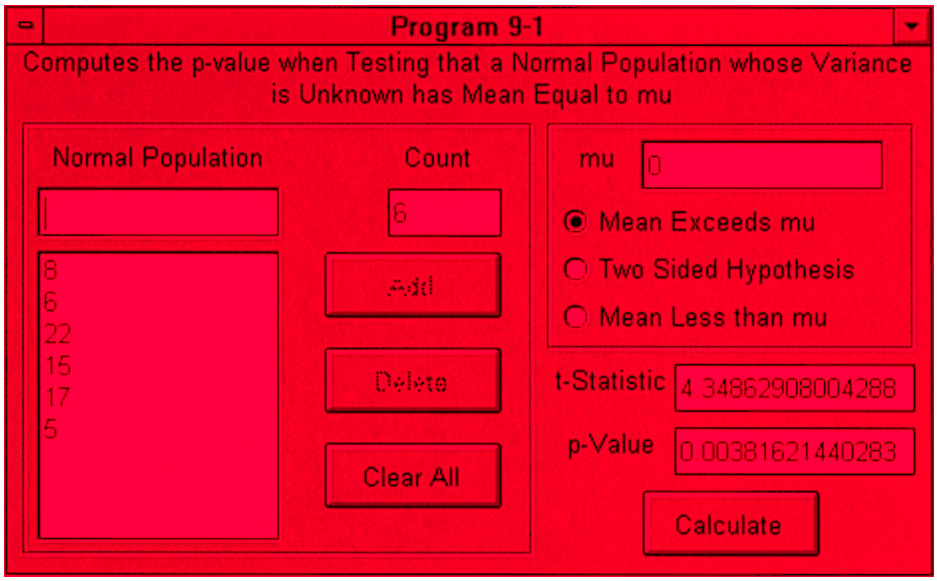
Ejemplo 10.9 El gerente de una cadena de grandes almacenes pretende averiguar si los anuncios tienden a incrementar sus ventas de zapatos de mujer. Para ello, observó el número de pares vendidos en seis establecimientos durante un periodo de dos semanas. Durante la primera semana no se emitieron anuncios, sino que comenzaron a emitirse al comienzo de la segunda semana. Si cualquier cambio en las ventas se debe únicamente a la campaña de anuncios, ¿prueban los datos que la campaña incrementa el número medio de ventas? Use un nivel de significación del 1%.

Establecimiento	**Ventas de la primera semana**	**Ventas de la segunda semana**
1	46	54
2	54	60
3	74	96
4	60	75
5	63	80
6	45	50

Solución Denotando por D_i los incrementos de ventas en el supermercado i, se quiere comprobar si los datos son suficientemente significativos para establecer que $\mu_d > 0$. Para ello, se necesita contrastar

$$H_0: \mu_d \leq 0 \qquad \text{frente a} \qquad H_1: \mu_d > 0$$

Con los valores 8, 6, 22, 15, 17, 5, se ha ejecutado el Programa 9-1, y se han obtenido los resultados siguientes:



Por tanto, la hipótesis de que la campaña de anuncios no implica incrementos en las ventas se ha de rechazar con cualquier nivel de significación mayor o igual a 0,0038. Por consiguiente, se rechaza al nivel de significación del 1%. ■

Problemas

1. Los siguientes datos se refieren a los ritmos cardiacos (en pulsaciones por minuto) de 12 individuos antes y después de consumir tabaco de mascar. Dichos sujetos eran usuarios regulares de esta sustancia.

Individuo	Ritmo cardiaco antes del consumo	Ritmo cardiaco después del consumo
1	73	77
2	67	69
3	68	73
4	60	70
5	76	74
6	80	88
7	73	76
8	77	82
9	66	69
10	58	61
11	82	84
12	78	80

   (a) Contraste la hipótesis, al nivel de significación del 5%, de que el ritmo cardiaco antes del consumo no varía con respecto al ritmo cardiaco posterior al consumo.

   (b) ¿Cuál es el p valor resultante?

2. Un vendedor de calzado mantiene que el uso de las zapatillas de carrera de su compañía permite hacer mejores tiempos. Para comprobar esta idea, un entrenador reunió a 10 velocistas, a los que dividió aleatoriamente en dos grupos de 5 individuos cada uno. Los miembros del primer grupo corrieron 100 yardas usando el calzado de carrera habitual, mientras que los del segundo grupo corrieron la misma distancia usando las zapatillas de la compañía. Tras un descanso, el grupo que había corrido con las zapatillas habituales las cambiaron por las de la compañía y los miembros del otro grupo se calzaron las zapatillas habituales. Tras ello, todos volvieron a correr 100 yardas. Resultaron los datos siguientes:

	Corredor									
	1	2	3	4	5	6	7	8	9	10
Tiempos (zapatillas habituales)	10,5	10,3	11,0	10,9	11,3	9,9	10,1	10,7	12,2	11,1
Tiempos (zapatillas nuevas)	10,3	10,0	10,6	11,1	11,0	9,8	10,2	10,5	11,8	10,5

¿Prueban estos datos la idea del vendedor de calzado de que las zapatillas nuevas de su compañía proporcionan, en media, menores tiempos? Use un nivel de significación del 10%. ¿Qué ocurriría si el nivel de significación fuera del 5%?

3. Utilice el contraste de la *t* con los siguientes datos apareados para contrastar

$$H_0: \mu_x = \mu_y \qquad \text{frente a} \qquad H_1: \mu_x \neq \mu_y$$

al nivel de significación del 5%.

	i										
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
X_i	122	132	141	127	141	119	124	131	145	140	135
Y_i	134	126	133	122	155	116	118	137	140	133	142

4. Una pregunta de interés médico es si correr moderadamente conduce a una reducción de la presión sanguínea sistólica. Para averiguarlo, ocho voluntarios no corredores convinieron en empezar un programa de carrera durante un mes. Al final del mes, se midieron sus presiones sanguíneas y se compararon con los valores anteriores; resultaron los siguientes datos:

	Individuo							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Presión sanguínea antes	134	122	118	130	144	125	127	133
Presión sanguínea después	130	120	123	127	138	121	132	135

   (a) Supongamos que se quiere ver si estos datos son suficientemente significativos para probar que el programa de carrera tiende a reducir la presión sanguínea sistólica. Determine las hipótesis nula y alternativa.

   (b) ¿Prueban los datos la hipótesis indicada en (a) al nivel de significación del 5%?

   (c) ¿Prueban los datos que la hipótesis es falsa?

   (d) ¿Cómo presentaría los resultados de este experimento a una persona médica no especializada en Estadística?

5. La siguiente tabla muestra las puntuaciones obtenidas en un test de inteligencia por 14 pares de mellizos monocigóticos (usualmente llamados *idénticos*) que vivieron en entornos separados tras su nacimiento. Un miembro de cada par se crió con al menos uno de sus progenitores biológicos, mientras que el otro creció en un hogar donde no habitaba ninguno de sus progenitores biológicos. El test IQ utilizado se conoce en la literatura de psicología como el test IQ de los "dominós".

Mellizo educado por su madre o su padre	Mellizo no educado por ninguno de sus progenitores	Mellizo educado por su madre o su padre	Mellizo no educado por ninguno de sus progenitores
23	18	22	15
30	25	31	23
25	28	29	27
18	22	24	26
19	14	28	19
25	34	31	30
28	36	27	28

(a) Contraste la hipótesis de que la puntuación media del test IQ de los mellizos no se ve afectada por el hecho de que fuera educado por uno de sus progenitores biológicos. Utilice un nivel de significación del 5%.

(b) ¿Qué conclusiones se pueden sacar, si es que es posible, del contraste empleado?

6. Considere el problema 2 de la sección 10.4. Suponga que las mismas mujeres fueron observadas en los dos meses y que los datos que aparecen en cada una de las columnas se refieren a la ingesta de grasa durante el verano y el invierno.

(a) Contraste la hipótesis de que no existe diferencia entre el consumo de grasa en verano e invierno. Use un nivel de significación del 5%.

(b) Repita el apartado (a) usando esta vez un nivel del 1%.

7. Los siguientes datos reflejan las puntuaciones obtenidas por 12 estudiantes universitarios en dos test IQ. Uno de los test se realizó antes de que los estudiantes siguieran un curso de Estadística, mientras que el otro se realizó después.

Estudiante	Puntuación IQ antes del curso	Puntuación IQ después del curso
1	104	111
2	125	120
3	127	138
4	102	113
5	140	142
6	122	130
7	118	114
8	110	121
9	126	135
10	138	145
11	116	118
12	125	125

Utilice estos datos para contrastar la hipótesis de que la puntuación de los estudiantes en el test IQ no tiende a ser diferente tras el seguimiento del curso. Use un nivel de significación del 5%.

8. Para ver si existen diferencias entre los salarios iniciales de los graduados masculinos y femeninos en Derecho se seleccionó un conjunto de ocho compañías. En cada una de ellas se eligió aleatoriamente a una mujer y un hombre entre el personal recientemente contratado. De la entrevista tenida con los elegidos se obtuvo la siguiente información:

	Compañía							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Salario de las mujeres	52	53,2	78	75	62,5	72	39	49
Salario de los hombres	54	55,5	78	81	64,5	70	42	51

Use estos datos para contrastar la hipótesis, al nivel de significación del 10%, de que los salarios iniciales son iguales para ambos sexos.

9. Para estudiar la efectividad de cierta dieta proteínica líquida comercializada, los servicios de la Administración muestrearon a nueve individuos a los que se sometió a un programa dietético. Se registraron sus pesos antes del programa y seis meses después del programa de dos semanas de duración al que fueron sometidos. Resultaron los siguientes datos:

Persona	**Peso antes**	**Peso después**
1	197	185
2	212	220
3	188	180
4	226	217
5	170	185
6	194	197
7	233	219
8	166	170
9	205	202

Supongamos que se quiere determinar si estos datos prueban que la dieta es efectiva, en el sentido de que la pérdida de peso esperada tras los seis meses es positiva.

(a) ¿Cuál es la hipótesis nula que se debería contrastar y cuál es la alternativa?

(b) ¿Prueban estos datos que la dieta funciona? Utilice un nivel del 5%.

10. Los siguientes datos muestran ciertas tasas de defunción de automovilistas por 100 millones de millas recorridas para una muestra aleatoria de Estados en los años 1985, 1989 y 2001.

Estado	**Tasa en 1985**	**Tasa en 1989**	**Tasa en 2001**
Arkansas	3,4	3,3	2,1
Colorado	2,4	1,9	1,7
Indiana	2,6	1,9	1,3
Kentucky	2,6	2,4	1,8
Massachusetts	1,9	1,7	0,9
Ohio	2,1	2,1	1,3
Tennessee	3,4	2,3	1,8
Wyoming	2,7	2,3	2,3

Fuente: Datos de accidentes, Consejo de Seguridad Nacional, Chicago.

(a) ¿Establecen estos datos que, al nivel de significación del 5%, las tasas de defunción eran inferiores en 1989 que en 1985?

(b) ¿Establecen estos datos, al nivel de significación del 5%, que la tasa de de función era menor en 2001?

(c) ¿Cuál es el p valor de los contrastes de los apartados (a) y (b)?

11. Los siguientes datos muestran las tasas de matrimonio por 1000 habitantes en una muestra aleatoria de países.

Tasas brutas de matrimonio para los países seleccionados (por 1000 habitantes)

País	1999	1998	1997	1990
Australia	6,0	—	5,8	6,9
Austria	4,8	4,8	5,1	5,8
Bélgica	4,3	4,4	4,7	6,6
Bulgaria	4,2	4,3	4,1	6,7
República Checa	5,2	5,4	5,6	8,4
Dinamarca	6,6	6,5	6,4	6,1
Finlandia	4,7	4,5	4,6	4,8
Alemania	5,2	5,1	5,2	6,5
Grecia	6,4	5,5	5,7	5,8
Hungría	4,5	4,5	4,6	6,4
Irlanda	4,9	—	4,3	5,0
Israel	5,9	—	5,6	7,0
Japón	6,3	6,3	6,2	5,8
Luxemburgo	4,9	—	4,8	6,2
Holanda	5,6	—	5,5	6,4
Nueva Zelanda	5,3	—	5,3	7,0
Noruega	5,3	—	—	5,2
Polonia	5,7	5,4	5,3	6,7
Portugal	6,8	6,7	6,5	7,3
Rumanía	6,5	6,4	6,5	8,3
Rusia	5,8	5,8	6,3	8,9
Suecia	4,0	3,5	3,7	4,7
Suiza	4,9	—	5,3	6,9

Contraste la hipótesis de que las tasas de matrimonio en 1999 eran superiores a las de 1990.

10.6 Contraste de igualdad de proporciones poblacionales

Consideremos dos poblaciones grandes y denotemos por p_1 y p_2, respectivamente, las proporciones de miembros de estas dos poblaciones que poseen cierta característica de inte-

rés. Supongamos que estamos interesados en contrastar la hipótesis de que estas proporciones son iguales frente a la alternativa de que son distintas. Esto es, pretendemos contrastar

$$H_0: p_1 = p_2 \qquad \text{frente a} \qquad H_1: p_1 \neq p_2$$

Para contrastar esta hipótesis nula supongamos que hemos extraído de estas poblaciones dos muestras independientes con tamaños respectivos n_1 y n_2. Representemos por X_1 y X_2 el número de elementos en cada una de estas dos muestras que poseen la característica.

Denotemos por $\hat{p}_1$ y $\hat{p}_2$ las proporciones de miembros de las dos muestras que presentan la característica. Esto es, $\hat{p}_1 = X_1/n_1$ y $\hat{p}_2 = X_2/n_2$. Puesto que $\hat{p}_1$ y $\hat{p}_2$ son los respectivos estimadores de p_1 y p_2, es evidente que se debe rechazar H_0 cuando $\hat{p}_1$ y $\hat{p}_2$ sean muy diferentes; esto es, cuando $|\hat{p}_1 - \hat{p}_2|$ sea suficientemente grande. Para ver lo diferentes que deben ser $\hat{p}_1$ y $\hat{p}_2$ para justificar el rechazo de H_0, primero se necesita determinar la distribución de probabilidad de $\hat{p}_1 - \hat{p}_2$.

Recordemos de la sección 7.5 que la media y la varianza de la proporción de miembros de la primera muestra que tienen la característica vienen dadas por

$$E[\hat{p}_1] = p_1 \qquad \text{Var}(\hat{p}_1) = \frac{p_1(1 - p_1)}{n_1}$$

y, de igual forma, para la segunda muestra,

$$E[\hat{p}_2] = p_2 \qquad \text{Var}(\hat{p}_2) = \frac{p_2(1 - p_2)}{n_2}$$

Por consiguiente, se ve que:

$$\begin{aligned} E[\hat{p}_1 - \hat{p}_2] &= E[\hat{p}_1] - E[\hat{p}_2] \\ &= p_1 - p_2 \end{aligned}$$

$$\begin{aligned} \text{Var}(\hat{p}_1 - \hat{p}_2) &= \text{Var}(\hat{p}_1) + \text{Var}(\hat{p}_2) \\ &= \frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2} \end{aligned}$$

Adicionalmente, si se supone que n_1 y n_2 son razonablemente grandes $\hat{p}_1$ y $\hat{p}_2$ seguirán una distribución aproximadamente normal y, por consiguiente, también lo hará su diferencia $\hat{p}_1 - \hat{p}_2$. Como resultado, la variable estandarizada

$$\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{p_1(1 - p_1)/n_1 + p_2(1 - p_2)/n_2}}$$

seguirá aproximadamente una distribución igual a la de una variable aleatoria normal estándar.

Supongamos ahora que H_0 es cierta y que, por tanto, las proporciones son iguales. Denotemos por p ambos valores iguales; esto es, $p_1 = p_2 = p$. En este caso, $p_1 - p_2 = 0$ y, por tanto, el valor de

$$W = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p(1 - p)/n_1 + p(1 - p)/n_2}} \tag{10.3}$$

seguirá aproximadamente una distribución normal estándar. Sin embargo, no podemos basar directamente nuestro contraste en W, puesto que depende del valor desconocido p. Sin embargo, se puede estimar p si se observa que en la muestra combinada de tamaño $n_1 + n_2$ existe un total de $X_1 + X_2 = n_1\hat{p}_1 + n_2\hat{p}_2$ elementos que tienen la característica de interés. Por consiguiente, cuando H_0 es cierta y cada una de las poblaciones tiene la misma proporción de miembros con la característica, el estimador natural de la proporción p común es el siguiente:

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{X_1 + X_2}{n_1 + n_2}$$

El estimador $\hat{p}$ se llama estimador *combinado* (“*pooled*”) de p.

Sustituiremos ahora el estimador $\hat{p}$ en lugar del parámetro desconocido p en la ecuación (10.3) de W, y basaremos nuestro contraste en la expresión resultante. Esto es, usaremos como estadístico del contraste

$$\text{TS} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})/n_1 + \hat{p}(1 - \hat{p})/n_2}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{(1/n_1 + 1/n_2)\,\hat{p}(1 - \hat{p})}}$$

Se puede demostrar que para valores razonablemente grandes de n_1 y n_2 (es suficiente que ambos sean superiores a 30), TS sigue, cuando H_0 es cierta, una distribución que es aproximadamente igual a una normal estándar. Así pues, el contraste de

$$H_0: p_1 = p_2 \qquad \text{frente a} \qquad H_1: p_1 \neq p_2$$

al nivel de significación α consistirá en

Rechazar H_0	si $\lvert\text{TS}\rvert \geq z_{\alpha/2}$
No rechazar H_0	en otro caso

El contraste también se puede realizar si primero se determina el valor del estadístico del contraste, digamos que este es igual a v, y después se calcula el p valor dado por

$$p \text{ valor} = P\{|Z| \geq |v|\} = 2P\{Z \geq |v|\}$$

donde Z es (como siempre) una variable aleatoria normal estándar.

Ejemplo 10.10 En los procedimientos criminales, en ocasiones el juez envía al acusado a prisión, y, otras veces, no. Una pregunta que surgió en los círculos legales es si la decisión del juez se ve afectada por (1) si el acusado se declaró culpable o (2) si se declaró inocente, pero después se probó que era culpable. Los siguientes datos se refieren a individuos, enviados a prisión, acusados de robos de segundo nivel.

74, de 142 que se habían declarado culpables, fueron a prisión
61, de 72 que se habían declarado no culpables, fueron a prisión

¿Estos datos indican que la posibilidad de que un acusado sea enviado a prisión depende de que se haya delarado culpable?

Solución Denotemos por p_1 la probabilidad de que un acusado que se ha declarado culpable sea enviado a prisión, y por p_2 la probabilidad equivalente para un acusado que se ha declarado inocente pero que después se haya probado que era culpable. Para ver si los datos son suficientemente significativos para probar que $p_1 \neq p_2$, se necesita contrastar

$$H_0: p_1 = p_2 \qquad \text{frente a} \qquad H_1: p_1 \neq p_2$$

Los datos indican que:

$$n_1 = 142 \qquad \hat{p}_1 = \frac{74}{142} = 0{,}5211$$

$$n_2 = 72 \qquad \hat{p}_2 = \frac{61}{72} = 0{,}8472$$

El valor del estimador combinado $\hat{p}$ es

$$\hat{p} = \frac{74 + 61}{142 + 72} = 0{,}6308$$

y el valor del estadístico del contraste es

$$\text{TS} = \frac{0{,}5211 - 0{,}8472}{\sqrt{(1/142 + 1/72)(0{,}6308)(1 - 0{,}6308)}} = -4{,}67$$

El p valor viene dado por

$$p \text{ valor} = 2P\{Z \geq 4.67\} \approx 0$$

Para este p valor tan pequeño la hipótesis nula debe ser rechazada. Es decir, se puede concluir que la decisión judicial, con respecto a los acusados enviados a prisión, se ve efectivamente afectada por el hecho de que el acusado se haya declarado culpable o inocente. (No se puede, sin embargo, concluir que declararse culpable es una buena estrategia para el acusado en lo referido a evitar la prisión. La razón por la que no se puede es que un acusado que se declara inocente tiene posibilidad de ser absuelto.) ■

Nuestro siguiente ejemplo ilustra las dificultades que presenta la modelización de modelos de fenómenos reales.

Ejemplo 10.11: Predicción del sexo de un hijo Supongamos que estamos interesados en determinar un modelo para predecir el sexo de los futuros hijos de las familias. El modelo más sencillo podría partir del supuesto de que, con independencia de la situación de la familia que se trate, cada nuevo nacimiento tiene una probabilidad p_0 de ser varón. (Es interesante resaltar que los datos existentes indican que p_0 se encuentra más próximo a 0,51 que a 0,50.)

En cierto sentido es sorprendente que este sencillo modelo no se mantiene cuando se analizan los datos reales. Por ejemplo, Malinvaud presentó en 1955 los datos sobre el sexo de los miembros de las familias francesas. Con respecto a las familias con cuatro o más hijos, Malinvaud encontró que, en 36 694 de ellas, los tres primeros nacimientos (es decir, los tres descendientes de mayor edad) habían sido hembras, mientras que existían 42 212 de tales familias cuyos tres primeros nacimientos habían sido todos varones. Los datos de Malinvaud indicaban que, en aquellas familias cuyos tres primeros nacimientos fueron todos hijas, el siguiente nacimiento era varón en el 46,9% de los casos; por el contrario, en las familias cuyos tres primeros nacimientos fueron varones, el siguiente nacimiento era varón en el 52,3% de los casos.

Denotemos por p_1 la probabilidad de que el siguiente nacimiento sea un varón en las familias que tienen tres hijas en la actualidad, y denotemos por p_2 la probabilidad equivalente para las familias que tienen actualmente tres hijos varones. Si utilizamos los datos de Malinvaud para contrastar

$$H_0: p_1 = p_2 \qquad \text{frente a} \qquad H_1: p_1 \neq p_2$$

se tiene:

$$n_1 = 36{,}694 \qquad n_2 = 42{,}212$$
$$\hat{p}_1 = 0{,}496 \qquad \hat{p}_2 = 0{,}523$$

de donde

$$\hat{p} = \frac{36\,694(0{,}496) + 42\,212(0{,}523)}{36\,694 + 42\,212} = 0{,}51044$$

Por consiguiente, el valor del estadístico del contraste es

$$TS = \frac{0{,}496 - 0{,}523}{\sqrt{(1/36\,694 + 1/42\,212)(0{,}5104)(1 - 0{,}5104)}} = -7{,}567$$

Puesto que $|TS| \geq z_{0{,}005} = 2{,}58$, la hipótesis nula (esto es, que la probabilidad de que el siguiente nacimiento sea un varón es la misma para las familias que tienen en la actualidad tres hijas o tres hijos) debe rechazarse al nivel de significación del 1%. De hecho, el p valor de estos datos es

$$p \text{ valor} = P\{|Z| \geq 7{,}567\} = 2P\{Z \geq 7{,}567\} \approx 0$$

Esto muestra que cualquier modelo que asuma que la probabilidad de que el sexo de un nacimiento no depende de la situación presente de la familia no es consistente con los datos

existentes. (Un modelo que sí que es consistente con los datos mostrados es suponer que cada familia tiene su propia probabilidad de que un nuevo nacimiento sea un varón, manteniéndose esta probabilidad constante con independencia de la situación actual de la familia. Dicha probabilidad, sin embargo, difiere de familia a familia.) ■

La forma ideal de contrastar la hipótesis de que los resultados de dos tratamientos son idénticos es dividir aleatoriamente a un grupo de personas en dos partes, una de ellas recibirá el primer tratamiento y la otra recibirá el segundo. Sin embargo, no siempre es posible realizar dicha división aleatoria. Por ejemplo, si se desea estudiar si el consumo de alcohol incrementa el riesgo de cáncer de próstata, no se puede obligar a los componentes de una muestra aleatoria a beber alcohol. Una alternativa para analizar la hipótesis consiste en llevar a cabo un estudio *observacional* que comienza con una elección aleatoria de un conjunto de bebedores y otro de no bebedores. Estos conjuntos son observados durante un periodo de tiempo, y después se usan los datos resultantes para contrastar la hipótesis de que los miembros de los dos grupos tienen el mismo riesgo de padecer cáncer de próstata.

Nuestro siguiente ejemplo muestra otra forma de realizar un estudio observacional.

Ejemplo 10.12 En 1970, los investigadores Herbst, Ulfelder y Poskancer (H-U-P) sospecharon que la causa del cáncer vaginal en mujeres jóvenes, una enfermedad bastante rara, podía deberse al consumo materno de dietilestibestrol (medicamento normalmente denotado como DES) durante el embarazo. Para estudiar esta posibilidad, los investigadores podrían haber hecho un estudio observacional mediante un grupo (de tratamiento) de mujeres cuyas madres tomaron DES durante el embarazo, y un grupo (de control) de mujeres cuyas madres no lo tomaron. Después se podrían haber observado ambos grupos durante cierto periodo de tiempo y los datos resultantes se podrían utilizar para contrastar la hipótesis de que las probabilidades de contraer un cáncer vaginal eran iguales en ambos grupos. Sin embargo, debido a que el cáncer vaginal es tan raro (en los dos grupos), tal estudio hubiera requerido un gran número de individuos de los dos grupos y probablemente también un gran número de años de de seguimiento para poder obtener resultados significativos. Por consiguiente, H-U-P optaron por un tipo de estudio observacional distinto. Encontraron a un grupo de 8 mujeres con una edad comprendida entre 15 y 22 años que padecían cáncer vaginal. A cada una de estas mujeres (llamadas casos) se las asignaron otras 4 mujeres, llamadas *referentes* o *controles*. Ninguno de los referentes de un caso padecía cáncer y todos habían nacido en el mismo hospital, con menos de 5 días de diferencia y en el mismo tipo de habitación (bien privada o pública) que el caso. Razonando que, si el DES no tenía efecto sobre el cáncer vaginal, la probabilidad, digamos p_c, de que las madres de los casos tomaran DES debía ser igual a la probabilidad, digamos p_r, de que las madres de los referentes tomaran DES, los investigadores H-U-P decidieron contrastar

$$H_0: p_c = p_r \qquad \text{frente a} \qquad H_1: p_c \neq p_r$$

Al descubrir que 7 de los 8 casos tenían madres que tomaron DES durante el embarazo mientras que ninguna de las madres de los referentes lo había tomado, los investigadores concluyeron que existía una fuerte asociación entre el DES y el cáncer vaginal (véase Herbst, A., Ulfelder, H. y Poskancer, D., "Adenocarcinoma of the Vagina: Association of Maternal Stilbestrol Therapy with Tumor Appearance in Young Women". *New England*

Journal of Medicine, **284**, 878-881, 1971). (El p valor para estos datos es aproximadamente 0.) ■

Si se deseara verificar la hipótesis unilateral de que p_1 es mayor que p_2, este hecho se debería tomar como hipótesis alternativa y, por tanto, se tendría que contrastar

$$H_0: p_1 \leq p_2 \qquad \text{frente a} \qquad H_1: p_1 > p_2$$

Se ha de emplear el mismo estadístico del contraste, TS, que antes, pero ahora se deberá rechazar H_0 sólo cuando TS sea grande (puesto que esto ocurre cuando $\hat{p}_1 - \hat{p}_2$ es grande). Por consiguiente, el contraste unilateral a nivel de significación α consistirá en

Rechazar H_0	si TS $\geq z_\alpha$
No rechazar H_0	en otro caso

Alternativamente, si el valor del estadístico, TS, del contraste es v, el p valor resultante es

$$p \text{ valor} = P\{Z \geq v\}$$

donde Z es una normal estándar.

Observación: El contraste de

$$H_0: p_1 \leq p_2 \qquad \text{frente a} \qquad H_1: p_1 > p_2$$

es el mismo que el de

$$H_0: p_1 = p_2 \qquad \text{frente a} \qquad H_1: p_1 > p_2$$

Esto ocurre porque en ambos casos se debe rechazar la hipótesis cuando $\hat{p}_1 - \hat{p}_2$ sea tan grande como para que fuera muy improbable haber obtenido un valor tan alto, si no ocurriera que p_1 fuera mayor que p_2.

Ejemplo 10.13 Un productor ha diseñado un nuevo método para producir chips de ordenador. Intuye que este nuevo método reduce la proporción de chips que resultan ser defectuosos. Para verificar esta intuición, se produjeron 320 chips por el nuevo método y 360 por el antiguo. El resultado fue que 76 de los primeros y 94 de los segundos resultaron defectuosos. ¿Proporciona esto una evidencia significativa para que el productor concluya que con el nuevo método se conseguirá una proporción menor de chips defectuosos? Utilice un nivel de significación del 5%.

Solución Denotemos por p_1 la probabilidad de que uno de los chips producidos por el método antiguo sea defectuoso, y por p_2 la correspondiente probabilidad para los chips producidos por el nuevo método. Para concluir que $p_1 > p_2$ se precisaría rechazar H_0 al contrastar

$$H_0: p_1 \leq p_2 \qquad \text{frente a} \qquad H_1: p_1 > p_2$$

Los datos son:

$$n_1 = 360 \qquad n_2 = 320$$
$$\hat{p}_1 = \frac{94}{360} = 0{,}2611 \qquad \hat{p}_2 = \frac{76}{320} = 0{,}2375$$

El valor del estimador combinado ("*pooled*") es pues

$$\hat{p} = \frac{94 + 76}{360 + 320} = 0{,}25$$

De donde, el valor del estadístico del contraste es

$$\text{TS} = \frac{0{,}2611 - 0{,}2375}{\sqrt{(1/360 + 1/320)(0{,}25)(0{,}75)}} = 0{,}7094$$

Puesto que $z_{0{,}05} = 1{,}645$, no se puede rechazar la hipótesis nula al nivel de significación del 5%. Esto es, la evidencia no es suficientemente significativa para concluir que con el nuevo método se conseguirá un menor porcentaje de chips defectuosos que con el método antiguo.

El p valor para estos datos es

$$p \text{ valor} = P\{Z \geq 0{,}7094\} = 0{,}239$$

lo que indica que, si se asume que las dos probabilidades son iguales, en 24 de cada 100 casos se podrían obtener valores de TS al menos tan grandes como el observado. ■

La tabla 10.5 detalla los contrastes considerados en esta sección.

Tabla 10.5 Contrastes relativos a dos probabilidades binomiales.

Las proporciones de miembros de dos poblaciones que poseen cierta característica son p_1 y p_2. Se elige una muestra aleatoria de tamaño n_1 procedente de la primera población, e independientemente de la anterior se elige igualmente una muestra aleatoria de tamaño n_2 procedente de la segunda población. El número de elementos muestrales que poseen la característica en cada una de las dos muestras serán X_1 y X_2, respectivamente.

$$\hat{p}_1 = \frac{X_1}{n_1} \qquad \hat{p}_2 = \frac{X_2}{n_2}$$

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

H_0	H_1	Estadístico, TS, del contraste	Contraste a nivel de significación α	p valor si TS $= v$
$p_1 = p_2$	$p_1 \neq p_2$	$\dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{(1/n_1 + 1/n_2)\hat{p}(1 - \hat{p})}}$	Rechazar H_0 si $\lvert\text{TS}\rvert \geq z_{\alpha/2}$ No rechazar en otro caso	$2P\{Z \geq \lvert v\rvert\}$
$p_1 \leq p_2$	$p_1 > p_2$	$\dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{(1/n_1 + 1/n_2)\hat{p}(1 - \hat{p})}}$	Rechazar H_0 si $\text{TS} \geq z_{\alpha}$ No rechazar en otro caso	$P\{Z \geq v\}$

La Estadística en perspectiva

No se malinterprete un rechazo

Uno debe ser cuidadoso cuando se está analizando lo que realmente significa el rechazo de la hipótesis nula, porque en ocasiones se dan interpretaciones no garantizadas realmente por los datos disponibles. Por ejemplo, supongamos que se lleva a cabo un contraste de hipótesis para estudiar si las probabilidades de que un paciente sobreviva a una determinada operación son iguales en los hospitales A y B. Supongamos que una muestra aleatoria de operaciones realizadas en el hospital A indica que 72 de 480 pacientes operados no han sobrevivido, mientras que una muestra extraída del hospital B indica que 30 de 360 no sobrevivieron. Aunque claramente se puede concluir a partir de estos datos que las probabilidades de supervivencia son desiguales, no se puede concluir que el hospital A no realiza su trabajo tan bien como el B, porque sin datos adicionales no se puede analizar si el hospital A está realizando más operaciones de alto riesgo que el B y que ésta es la razón por la que tiene una menor tasa de supervivencia.

Como ejemplo adicional que indica lo cuidadosos que debemos ser a la hora de interpretar el significado de rechazar una hipótesis, consideremos un estudio hipotético de los salarios de los vendedores varones y hembras de una gran compañía. Supongamos que una muestra de empleados, formada por 50 varones y 50 hembras, indicó que el salario de los varones es de 40 000 dólares mientras que el de las mujeres es de 36 000. Si las varianzas muestrales fueran pequeñas, el contraste de la hipótesis de que los salarios medios son iguales en las dos poblaciones conduciría a rechazar esta hipótesis. Pero, ¿qué se puede concluir de esto? Por ejemplo, ¿estaría justificado concluir que las mujeres están siendo discriminadas a la contra? La respuesta es que no se puede llegar a tal conclusión con la información presentada, porque pueden existir muchas justificaciones posibles para las aparentes diferencias en los salarios medios.

Una posibilidad podría ser que la mixtura de trabajadores con experiencia y sin experiencia es diferente en ambos sexos. Por ejemplo, si se considera el hecho de si un empleado ha trabajado más o menos de 5 años podría haber reflejado los datos siguientes:

Años de empleo	Número	Salario medio ($)
Hombres:		
Menos de 5	10	34 000
Más de 5	40	41 500
Total	50	40 000
Mujeres:		
Menos de 5	40	34 500
Más de 5	10	42 000
Total	50	36 000

Un total de 10 sobre 50 mujeres, por ejemplo, han estado contratadas más de 5 años, y su salario medio es de 42 000 dólares por año. Por consiguiente, vemos que, aunque el salario medio de los hombres es mayor que el de las mujeres, cuando se tiene en cuenta el tiempo de contratación, los empleados hembras están realmente recibiendo salarios mayores que sus equivalentes empleados varones.

Problemas

1. Se han propuesto dos métodos para producir transistores. Si con el método 1 se produjeron 20 transistores defectuosos de un total de 100 producidos, y con el método 2 se produjeron 12 transistores defectuosos igualmente de un total de 100 producidos, ¿se puede concluir que las proporciones de los transistores defectuosos que se producen con ambos métodos son diferentes?

   (a) Use un nivel de significación del 5%.

   (b) ¿Qué ocurriría si el nivel de significación fuera del 10%?

2. Se ha entrevistado a una muestra aleatoria de bebedores de café compuesta por 220 mujeres y 210 hombres. El resultado fue que 71 mujeres y 58 hombres manifestaron su preferencia por el café descafeinado. ¿Establecen estos datos, al nivel de significación del 5%, que la proporción de mujeres bebedoras de café que prefieren café descafeinado difiere de la correspondiente proporción de hombres? ¿Cuál es el p valor?

3. Una compañía de seguros de automóviles seleccionó dos muestras aleatorias de 300 suscriptores varones solteros y de 300 suscriptores varones casados, todos en edades comprendidas entre 25 y 30 años. Se registró el número de ellos que habían tenido algún tipo de accidente en los tres últimos años. Los datos resultantes fueron que un 19% de los solteros y un 12% de los casados habían tenido algún tipo de accidente.

   (a) ¿Establece esto que, al nivel de significación del 10%, existen diferencias entres las dos clases de suscriptores?

   (b) ¿Cuál es el p valor para el contraste del apartado (a)?

4. En 1976 se institucionalizó un extenso programa de vacunación contra la gripe. Aproximadamente 50 millones de los 220 millones de norteamericanos habían sido vacunados. De un grupo de 383 personas que padecieron la gripe, 202 habían sido vacunados.

   (a) Contraste la hipótesis, al nivel del 5%, de que la probabilidad de contraer la gripe es idéntica para la población de vacunados que para la de no vacunados.

   (b) ¿Indican los resultados del apartado (a) que la vacuna misma fue la causante de la gripe? ¿Puede imaginar otra explicación posible?

5. Se quieren comparar dos tipos de insecticidas. Para ello se fumigaron dos habitaciones de igual tamaño, una con el insecticida 1 y la otra con el insecticida 2. Después, se soltaron 100 insectos en cada habitación y, transcurridas 2 horas, se contó el número de insectos muertos. Suponga que el número de insectos muertos fue de 64 en la habitación fumigada con el insecticida 1, y de 52 en la otra habitación.

   (a) ¿Existe una evidencia suficientemente significativa para que rechacemos la hipótesis de que los dos insecticidas tienen la misma capacidad de eliminar insectos a un nivel del 5%?

   (b) ¿Cuál es el p valor para el contraste del apartado (a)?

6. Se seleccionaron dos muestras aleatorias de 100 residentes de San Francisco y de 100 de Los Ángeles, a los que se les preguntó sobre si estaban a favor de que se aumentara la edad para poder obtener el permiso de conducir. El resultado fue que 56 de los entrevistados de San Francisco y 45 de los de Los Ángeles se mostraron a favor.

   (a) ¿Son estos datos lo suficientemente fuertes como para establecer, al nivel de significación del 10%, que las proporciones de población a favor de la medida son diferentes en ambas ciudades?

   (b) ¿Qué ocurriría si el nivel fuera del 5%?

7. En 1983 una muestra aleatoria de 1000 científicos incluía a 212 mujeres. Por otra parte, una muestra aleatoria de 1000 científicos extraída en 1990 incluía a 272 mujeres. Utilice estos datos para contrastar la hipótesis de que las proporciones de mujeres eran las mismas entre los científicos de 1983 y 1990. Encuentre también el p valor.

8. En el ejemplo 10.11 se consideró un modelo para predecir el sexo de los descendientes. Una generalización del citado modelo podría ser suponer que el sexo en un nacimiento depende sólo del número de hijos previos de la familia y del número de varones en ellos. Si fuera así, el sexo del tercer hijo en familias cuyos hijos fueran un varón y una hembra no dependería de que el orden de los dos primeros nacimientos fuera chico-chica o chica-chico. Los siguientes datos muestran el sexo del tercer hijo en familias del tipo indicado con un primer descendiente varón o hembra. Esto permite distinguir si el primogénito fue chico o chica. (Por ejemplo, chico-chica significa que el primer descendiente fue un chico.)

Familias con chico-chica	**Familias con chica-chico**
412 chicos	560 chicos
418 chicas	544 chicas

   Use los datos dados para contrastar la hipótesis de que el sexo del tercer hijo en las familias que cuentan actualmente con un hijo y una hija no depende del orden por sexo de los dos primeros nacimientos. Utilice un nivel de significación del 5%.

9. De acuerdo con el Centro Nacional de Estadística Sanitaria, hubo un total de 330 535 mujeres afroamericanas y de 341 441 varones afroamericanos nacidos en 1988. En dicho año, se produjeron 1 483 487 nacimientos de chicas blancas y 1 562 675 nacimientos de chicos blancos. Utilice estos datos para contrastar la hipótesis de que la proporción de bebés afroamericanos hembras es igual a la de bebés blancos hembras. Use un nivel del significación del 5%. Calcule también el p valor.

10. Supongamos que una muestra de 480 operaciones de bypass de corazón en el hospital A mostró que 72 pacientes no habían sobrevivido, mientras que de una muestra de 360 operaciones en el hospital B no sobrevivieron 30 pacientes. Encuentre el p valor para el contraste de la hipótesis de que las probabilidades de supervivencia son iguales en los dos hospitales.

11. Recientemente, la Universidad de California, en un seminario de paternidad, añadió una conferencia sobre la importancia de utilizar asientos para niños en los automóviles. Esta decisión se tomó tras haber estudiado los resultados de un experimento en el que la conferencia se impartió a un grupo de partícipes en el seminario y no se impartió a otro grupo. En una entrevista de seguimiento, llevada a cabo un año más tarde, se encuestó a 82 parejas que habían asistido a la conferencia y a 120 que no. Entre las que habían asistido, 78 parejas manifestaron que siempre utilizaban asientos de niños en el coche, mientras que 90 de las parejas que no atendieron la conferencia lo indicaron igualmente.

    (a) Si se asume la exactitud de la información suministrada, ¿la diferencia es suficientemente significativa para concluir que la conferencia ha producido un incremento en el uso de asientos para niños en los automóviles? Use un nivel de significación del 5%.

    (b) ¿Cuál es el p valor?

12. En una encuesta reciente, 52 de 200 personas muestreadas manifestaron tener un arma en casa. En una encuesta anterior, 28 de 150 personas indicaron lo mismo. ¿Prueba esto que existe ahora más gente que tiene, o manifiesta tener, armas en casa?

    (a) Use un nivel de significación del 5%.

    (b) ¿Cuál es el p valor?

13. Para ver lo efectiva que es una nueva vacuna contra el resfriado común, se dividió a 204 trabajadores de una estación de esquí en dos grupos, de 102 cada uno. A los miembros del primer grupo se les inyectó la vacuna, mientras que a los miembros del segundo grupo se les inyectó un placebo. A final de la temporada de invierno, resultó que 29 de los individuos que fueron vacunados sufrieron al menos un resfriado, mientras que 34 de los que recibieron el placebo también lo sufrieron. ¿Prueba esto, al nivel de significación del 5%, que la vacuna es efectiva para prevenir resfriados?

14. La Sociedad Americana contra el Cáncer muestreó recientemente a 2500 adultos y comprobó que 738 de ellos eran fumadores. Una encuesta similar realizada en 1986 a 2000 adultos reflejó un total de 640 fumadores. ¿Prueban estas cifras que la proporción de adultos que fuman ha descendido desde 1986?

    (a) Utilice un nivel de significación del 5%.

    (b) Use un nivel de significación del 1%.

15. En un estudio reciente realizado a 22 000 físicos varones, se les suministró a la mitad de ellos una dosis diaria de aspirina, mientras que a la otra mitad se les suministró un placebo. El estudio se continuó durante un periodo de seis años. Durante este periodo, 104 de los que tomaron aspirina y 180 de los que tomaron el placebo sufrieron ataques de corazón. ¿Indican estos resultados que tomar una dosis diaria de aspirina reduce el riesgo de sufrir un ataque de corazón? Especifique la hipótesis nula y el p valor resultante.

16. En la década de 1970, la Administración de Estados Unidos llevó a cabo un experimento para comparar la cirugía de bypass en la arteria coronaria con la terapia medicamentosa, como tratamientos alternativos de la enfermedad de la arteria coronaria. El experimento

involucró a 596 pacientes, de los cuales 286 fueron aleatoriamente asignados para recibir un tratamiento quirúrgico, mientras que a los restantes se les asignó la terapia medicamentosa. Un total de 252 de los sometidos a cirugía y un total de 270 de los sometidos a la terapia medicamentosa aún vivían después de 3 años del tratamiento. Utilice estos datos para contrastar la hipótesis de que las probabilidades de supervivencia son iguales.

Términos clave

Contrastes con dos muestras: Contrastes que tratan las relaciones entre los parámetros de dos poblaciones distintas.

Contrastes con muestras apareadas: Contrastes donde los datos consisten en pares de variables dependientes.

Resumen

I. Contraste de igualdad de medias poblacionales: muestras independientes

Supongamos que $X_1, \ldots, X_n$ y $Y_1, \ldots, Y_m$ son muestras independientes procedentes de poblaciones normales con parámetros respectivos μ_x, σ_x^2 y μ_y, σ_y^2.

Caso 1: σ_x^2 *y* σ_y^2 *son conocidas.*

Para contrastar

$$H_0: \mu_x = \mu_y \qquad \text{frente a} \qquad H_1: \mu_x \neq \mu_y$$

utilice el estadístico del contraste

$$TS = \frac{\overline{X} - \overline{Y}}{\sqrt{\sigma_x^2/n + \sigma_y^2/m}}$$

El contraste a nivel de significación α consiste en

$$\begin{array}{ll} \text{Rechaza } H_0 & \text{si } |TS| \geq z_{\alpha/2} \\ \text{No rechazar } H_0 & \text{en otro caso} \end{array}$$

Si el valor de TS es v, entonces

$$p \text{ valor} = P\{|Z| \geq |v|\} = 2P\{Z \geq |v|\}$$

donde Z es una variable aleatoria normal estándar.

El contraste a nivel de significación α de

$$H_0: \mu_x \leq \mu_y \qquad \text{frente a} \qquad H_1: \mu_x > \mu_y$$

utiliza el mismo estadístico del contraste. Se deberá

$$\begin{array}{ll} \text{Rechazar } H_0 & \text{si } TS \geq z_\alpha \\ \text{No rechazar } H_0 & \text{en otro caso} \end{array}$$

Si TS $= v$, el p valor es

$$p \text{ valor} = P\{Z \geq v\}$$

Caso 2: σ_x^2 y σ_y^2 *son desconocidas, y n y m son grandes.*

Para contrastar

$$H_0: \mu_x = \mu_y \quad \text{frente a} \quad H_1: \mu_x \neq \mu_y$$

o

$$H_0: \mu_x \leq \mu_y \quad \text{frente a} \quad H_1: \mu_x > \mu_y$$

utilice el estadístico del contraste

$$\text{TS} = \frac{\overline{X} - \overline{Y}}{\sqrt{S_x^2/n + S_y^2/m}}$$

donde S_x^2 y S_y^2 son las respectivas varianzas muestrales. El estadístico del contraste, el contraste a nivel de significación α, y el p valor son exactamente los mismos que en el caso 1.

Caso 3: Se asume que σ_x^2 y σ_y^2 *son desconocidos pero iguales.*

Para contrastar

$$H_0: \mu_x = \mu_y \quad \text{frente a} \quad H_1: \mu_x \neq \mu_y$$

utilice el estadístico del contraste

$$\text{TS} = \frac{\overline{X} - \overline{Y}}{\sqrt{S_p^2(1/n + 1/m)}}$$

donde S_p^2, conocido como *estimador combinado* (*"pooled"*) de la varianza común, viene dado por

$$S_p^2 = \frac{n-1}{n+m-2} S_x^2 + \frac{m-1}{n+m-2} S_y^2$$

El contraste a nivel de significación α consiste en

Rechazar H_0 si $|\text{TS}| \geq t_{n+m-2,\,\alpha/2}$

No rechazar H_0 en otro caso

Si TS $= v$, el p valor es

$$p \text{ valor} = 2P\{t_{n+m-2} \geq |v|\}$$

En lo anterior, T_{n+m-2} es una variable aleatoria t con $n + m - 2$ grados de libertad, y $t_{n+m-2,\alpha}$ es tal que

$$P\{T_{n+m-2} \geq t_{n+m-2,\alpha}\} = \alpha$$

Para contrastar

$$H_0: \mu_x \leq \mu_y \quad \text{frente a} \quad H_1: \mu_x > \mu_y$$

utilice el mismo estadístico del contraste. El contraste a nivel de significación α actúa como sigue:

$$\begin{array}{ll} \text{Rechazar } H_0 & \text{si } TS \geq t_{n+m-2,\,\alpha} \\ \text{No rechazar } H_0 & \text{en otro caso} \end{array}$$

Si $TS = v$,

$$p \text{ valor} = P\{T_{n+m-2} \geq v\}$$

II. Contraste de igualdad de medias poblacionales: muestras apareadas. Supongamos que $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_n$ son muestras procedentes de poblaciones normales con medias μ_x y μ_y. Supongamos además que estas muestras no son independientes sino que los n pares de variables aleatorias X_i y Y_i son dependientes, $i = 1, \ldots, n$. Sea, para cada i,

$$D_i = X_i - Y_i$$

y supongamos que $D_1, \ldots, D_n$ constituyen una muestra procedente de una población normal. Hagamos

$$\mu_d = E[D_i] = \mu_x - \mu_y$$

Para contrastar

$$H_0: \mu_x = \mu_y \quad \text{frente a} \quad H_1: \mu_x \neq \mu_y$$

contraste las hipótesis equivalentes

$$H_0: \mu_d = 0 \quad \text{frente a} \quad H_1: \mu_d \neq 0$$

Contrastar, pues, que las dos muestras tienen medias iguales es equivalente a contrastar que una población normal tiene media 0. Esta última hipótesis se contrasta utilizando el contraste de la t presentado en la sección 9.4. El estadístico del contraste es

$$TS = \sqrt{n}\,\frac{\overline{D}}{S_d}$$

y el contraste a nivel α consistirá en

$$\begin{array}{ll} \text{Rechazar } H_0 & \text{si } |TS| \geq t_{n\,-\,1,\,\alpha/2} \\ \text{No rechazar } H_0 & \text{en otro caso} \end{array}$$

Si $TS = v$, entonces

$$p \text{ valor} = 2P\{T_{n\,-\,1} \geq |v|\}$$

Para contrastar las hipótesis unilaterales

$$H_0: \mu_x \leq \mu_y \quad \text{frente a} \quad H_1: \mu_x > \mu_y$$

use el estadístico del contraste

$$\text{TS} = \sqrt{n}\frac{\overline{D}}{S_d}$$

El contraste a nivel de significación α consiste en

$$\begin{array}{ll} \text{Rechazar } H_0 & \text{si TS} \geq t_{n-1,\,\alpha} \\ \text{No rechazar } H_0 & \text{en otro caso} \end{array}$$

Si TS $= \nu$, se tiene que

$$p \text{ valor} = P\{T_{n-1} \geq \nu\}$$

III. Contraste de igualdad de proporciones poblacionales. Consideremos dos poblaciones grandes, en las que algunos de sus miembros poseen cierta característica. Denotemos por p_1 y p_2, respectivamente, las proporciones de miembros de la primera y de la segunda población que poseen la característica citada. Supongamos que se extrae una muestra de tamaño n_1 procedente de la población 1 y que se extrae otra de tamaño n_2 procedente de la población 2. Denotemos por X_1 y X_2, respectivamente, el número de elementos de cada una de estas muestras que poseen la característica.

Representemos por

$$\hat{p}_1 = \frac{X_1}{n_1} \qquad \text{y} \qquad \hat{p}_2 = \frac{X_2}{n_2}$$

las proporciones muestrales de individuos que presentan la característica; y por

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

la proporción de elementos que poseen la característica en la muestra combinada.

Para contrastar

$$H_0: p_1 = p_2 \qquad \text{frente a} \qquad H_1: p_1 \neq p_2$$

utilice el estadístico del contraste

$$\text{TS} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{(1/n_1 + 1/n_2)\hat{p}(1 - \hat{p})}}$$

El contraste a nivel de significación α consiste en

$$\begin{array}{ll} \text{Rechazar } H_0 & \text{si TS} \geq t_{\alpha/2} \\ \text{No rechazar } H_0 & \text{en otro caso} \end{array}$$

Si el valor de TS es ν, se cumple que

$$p \text{ valor} = 2P\{Z \geq |\nu|\}$$

Para contrastar

$$H_0: p_1 \leq p_2 \qquad \text{frente a} \qquad H_1: p_1 > p_2$$

utilice el estadístico del contraste

$$\begin{array}{ll} \text{Rechazar } H_0 & \text{si } TS \geq t_\alpha \\ \text{No rechazar } H_0 & \text{en otro caso} \end{array}$$

Si $TS = \nu$, el p valor viene dado por

$$p \text{ valor} = P\{Z \geq \nu\}$$

Observación: En lo anterior, al igual que en todo el texto, Z siempre representa una variable aleatoria normal estándar, y z_α es tal que

$$P\{Z \geq z_\alpha\} = \alpha$$

Problemas de repaso

1. Los siguientes datos representan los pesos de recién nacidos (en gramos) resultantes de un estudio que intentaba determinar el efecto en los fetos derivado del hecho de tener madres fumadoras durante el embarazo.

No fumadoras	Fumadoras
$n = 1820$	$m = 1340$
$\overline{X} = 3480$ gramos	$\overline{Y} = 3260$ gramos
$S_x = 9{,}2$ gramos	$S_y = 10{,}4$ gramos

   (a) Contraste la hipótesis, al nivel de significación del 5%, de que la media de los pesos de los recién nacidos es la misma con independencia de que la madre sea o no fumadora.

   (b) ¿Cuál es el p valor en el apartado (a)?

2. Se ha diseñado un estudio para comparar dos tratamientos reductores del riesgo de rechazo en los transplantes de corazón. El primer tratamiento consiste en suministrar al paciente salicilato sódico, y el segundo consiste en añadir a este medicamento un segundo medicamento, el azatioprine. El estudio fue realizado sobre ratas machos, utilizando un tipo de rata como donante y un segundo tipo como receptor. (El uso de tipos diferentes de ratas aseguraba que el receptor no sobreviría demasiado tiempo.) La variable de interés es el tiempo de supervivencia en días tras la recepción del corazón transplantado. Se obtuvieron los siguientes estadísticos sumariales.

Salicilato sódico	Salicilato sódico con azatioprine
$n = 14$	$m = 12$
$\overline{X} = 15{,}2$ días	$\overline{Y} = 14$ días
$S_x = 9{,}2$ días	$S_y = 9{,}0$ días

Contraste la hipótesis, al nivel de significación del 5%, de que los dos tratamientos son igualmente efectivos sobre la población de ratas.

3. Un estudio reciente relativo a los daños de rodilla en los jugadores de fútbol americano trataba de comparar dos tipos de zapatillas de fútbol. De un grupo de 1440 jugadores elegidos aleatoriamente, 240 utilizaron zapatillas multiclavo, mientras que los 1200 restantes usaron las zapatillas convencionales. Todos ellos jugaron sobre césped natural. Entre aquellos que usaron zapatillas multiclavo, 13 sufrieron daños en la rodilla; mientras que, entre los que utilizaron zapatillas convencionales, 78 sufrieron daños similares.
   (a) Contraste la hipótesis de que la probabilidad de sufrir daños de rodilla es la misma para ambos grupos de jugadores. Use un nivel de significación del 5%.
   (b) ¿Cuál es el p valor del apartado (a)?
   (c) ¿Los datos dados son suficientemente fuertes para establecer que las zapatillas multiclavo son mejores que las convencionales, en términos de reducir la probabilidad de sufrir daños de rodilla?
   (d) En el apartado (a), ¿para qué niveles de significación la evidencia es suficientemente fuerte?
4. Use los 60 primeros datos del Apéndice A. Contraste la hipótesis, al nivel de significación del 5%, de que las medias de hombres y mujeres son iguales en:
   (a) colesterol
   (b) presión sanguínea
5. Los siguientes datos proceden de un experimento realizado por Charles Darwin y publicado en su libro de 1876 *Los efectos de la fertilización cruzada y la autofertilización en el reino vegetal.* Los datos fueron inicialmente analizados por Francis Galton, primo de Darwin. Sin embargo, el análisis de Galton fue erróneo. El análisis correcto fue realmente realizado por R. A. Fisher.

   El experimento de Darwin se centró sobre 15 parejas de cierto tipo de planta de maiz, la variedad *Zea*. Una planta de cada par fue sometida a la fertilización cruzada, mientras que la otra planta fue autofertilizada. Las parejas crecieron en el mismo tiesto durante un tiempo, tras el cual se midieron sus alturas. Los datos fueron los siguientes:

Pareja	Planta con fertilización cruzada	Planta con autofertilización
1	23,5	17,375
2	12	20,375
3	21	20
4	22	20
5	19,125	18,375

(*Continúa*)

Pareja	Planta con fertilización cruzada	Planta con autofertilización
6	21,5	18,625
7	22,125	18,625
8	20,375	15,25
9	18,25	16,5
10	21,625	18
11	23,25	16,25
12	21	18
13	22,125	12,75
14	23	15,5
15	12	18

(a) Al nivel de significación del 5%, contraste la hipótesis de que la altura media de las plantas de maíz *Zea* con fertilización cruzada es igual a la de las plantas de maíz *Zea* autofertilizadas.

(b) Determine el p valor en el contraste de hipótesis del apartado (a).

6. Un debate que aún continúa en círculos de salud pública es el referente a los peligros de estar expuestos a la dioxina, un contaminante ambiental. Un estudio alemán publicado el 19 de octubre de 1991 en *The Lancet*, una revista médica británica, consideró registros de trabajadores de una compañía productora de un herbicida en la que se utilizaba la dioxina. Como grupo de control se tomaron trabajadores, con perfiles médicos similares al de los anteriores, de una compañía cercana de suministro de gas. Se obtuvieron los siguientes datos, relativos al número de trabajadores que habían muerto de cáncer.

	Grupo de control	Grupo expuesto a la dioxina
Tamaño muestral	1583	1242
Número de muertos de cáncer	113	123

(a) Contraste la hipótesis de que la probabilidad de morir de cáncer es idéntica para los dos grupos. Utilice un nivel de significación del 1%.

(b) Encuentre el p valor en el contraste del apartado (a).

7. Considere el problema 6. De los 1583 trabajadores de la compañía de gas cuyos registros fueron estudiados, hubo un total de 1184 hombres y 399 mujeres. De estos individuos, 93 hombres y 20 mujeres murieron de cáncer. Contraste la hipótesis, al nivel de significación del 5%, de que la probabilidad de morir de cáncer es idéntica para los trabajadores de ambos sexos.

8. Una muestra aleatoria de 56 mujeres reveló que 38 estaban a favor de un control de armas. Una muestra aleatoria de 64 hombres reveló que 32 de ellos se mostraban igualmente a favor. Utilice estos datos para contrastar la hipótesis de que las proporciones de hombres y de mujeres a favor del control de armas son iguales. Utilice un nivel de significación del 5%. ¿Cuál es el p valor?

9. Use los datos presentados en el problema de repaso 19 del capítulo 8 para contrastar la hipótesis de que la posibilidad de puntuar una carrera es la misma cuando hay un jugador fuera y uno en la segunda base que cuando no hay ningún jugador fuera y uno en la primera base.

10. Los siguientes datos se refieren a 100 partidos de béisbol profesional elegidos aleatoriamente y a 100 partidos de fútbol americano profesional en la temporada 1990-91. Los datos presentan, para los dos deportes, el número de partidos en los que el equipo que vencía al final de las tres cuartas partes del encuentro (final del séptimo *inning* en béisbol y final del tercer cuarto en fútbol) acabó perdiendo el partido.

Deporte	**Número de juegos**	**Número de juegos perdidos por el vencedor tras las primeras tres cuartas partes**
Béisbol	92	6
Fútbol	93	21

Encuentre el p valor del contraste de la hipótesis de que la probabilidad de perder el partido cuando se van ganando tras las tres primeras cuartas partes del encuentro es la misma en ambos deportes. (*Nota*: El número de partidos no es 100 porque 8 partidos de béisbol y 7 de los de fútbol estaban empatados tras las primeras tres cuartas partes.)

11. Los siguientes datos se refieren a las muestras de partidos del problema 10. Estos datos indican el número de partidos que ganó el equipo de casa.

Deporte	**Número de partidos**	**Número de partidos en los que ganó el equipo de casa**
Béisbol	100	53
Fútbol	99	57

Contraste la hipótesis, al nivel de significación del 5%, de que la proporción de partidos ganados por el equipo de casa es igual en ambos deportes.

12. Supongamos que en un contraste de H_0: $\mu_x = \mu_y$ frente a H_1: $\mu_x \neq \mu_y$ resulta rechazada H_0, al nivel de significación del 5%. ¿Cuál de las siguientes sentencias es (son) cierta(s)?

    (a) La diferencia entre las medias muestrales es significativa estadísticamente al nivel de significación del 1%.

    (b) La diferencia entre las medias muestrales es estadísticamente significativa al nivel de significación del 10%.

    (c) La diferencia entre las medias muestrales es igual a la diferencia entre las medias poblacionales.

13. Para verificar la hipótesis de que los niveles de plomo en la sangre son mayores en los niños cuyos padres trabajan en una factoría que utiliza plomo en su proceso de producción, los investigadores examinaron los niveles de plomo en la sangre de 33 niños cuyos padres tra-

bajaban en una fabrica de baterías. (Morton, D., Saah, A., Silberg, S., Owens, W., Roberts, M. y Saah, M., "Lead Absortion in Children of Employees in a Lead-Realated Industry", *American Journal of Epidemiology,* **115**, 549-555, 1982.) Cada uno de estos niños fue emparejado con otro niño que era de edad similar, vivía en un entorno parecido, estaba expuesto a un nivel de tráfico equivalente, pero cuyos padres no trabajaban con plomo. Se utilizaron después los niveles sanguíneos de los 33 casos (muestra 1) y de los 33 controles (muestra 2) para contrastar la hipótesis de que la media de los niveles sanguíneos de estos dos grupos coinciden. Si las resultantes medias muestrales y desviaciones típicas muestrales fueron

$$\overline{X}_1 = 0{,}015, \qquad S_1 = 0{,}004, \qquad \overline{X}_2 = 0{,}006, \qquad S_2 = 0{,}006$$

encuentre el p valor resultante. Asuma varianzas iguales.

14. Una científica que analiza el efecto de fumar sobre el corazón ha seleccionado aleatoriamente una muestra grande de fumadores y otra de no fumadores. Pretende estudiar estos dos grupos durante 5 años para ver si el número de ataques de corazón entre los miembros del grupo de fumadores es significativamente mayor que el número de ataques entre los no fumadores. Tal resultado, cree la científica, evidenciaría contundentemente la existencia de una asociación entre fumar y los ataques de corazón.
Si las afirmaciones siguientes fueran ciertas, ¿la creencia de la científica estaría justificada?

    1. La gente mayor presenta un mayor riesgo de ataque que la gente joven.
    2. Como grupo, los fumadores tienden de alguna forma a tener más edad que los no fumadores.

    Explique cómo se puede mejorar el diseño del experimento para que se pudieran extraer resultados más concluyentes.

11

Análisis de la varianza

La Estadística puede probar todo, incluso la verdad.

N. Moynihan (escritor inglés)

Se presenta un enfoque general, conocido como *análisis de la varianza* (ANOVA, de *Análisis of Variance*; en ocasiones denotada en castellano como ADEVA, por la misma razón), para hacer inferencias sobre las medias de varias variables aleatorias. En el ANOVA unifactorial, la media de cada variable depende sólo de un factor; a saber, la muestra a la que pertenece. En el ANOVA bifactorial, las variables aleatorias se pueden representar en una tabla de doble entrada, y la media de una variable depende tanto del factor fila como del factor columna. Se muestra cómo contrastar que la media de una variable aleatoria no depende de la fila en la que se encuentra, y también la hipótesis análoga de que la media no depende de la columna en la que se encuentra.

11.1 Introducción

En los años, en Estados Unidos, mucha gente ha expresado su temor a que cada vez en mayor medida una gran parte de su industria se está volviendo incapaz de competir de manera efectiva en la economía mundial. Por ejemplo, allí la opinión pública ha empezado a creer que los automóviles japoneses son de mejor calidad que sus equivalentes estadounidenses. Muchos consideran que Japón, y no Estados Unidos, es el líder mundial en la aplicación de las técnicas estadísticas de mejora de la calidad.

En las décadas de 1920 y 1930, los estadísticos industriales de Estados Unidos desarrollaron los métodos estadísticos de control de calidad. Estos métodos inicialmente analizaban la idoneidad de los procesos de producción existentes. En gran medida, se basaban en el uso de procedimientos estadísticos de muestreo que permitían que los estadísticos detectaran rápidamente cuándo algo había ido mal en los procesos de producción. Sin embargo, recientemente, el énfasis en el control estadístico de la calidad se ha trasladado desde la supervisión de un proceso de producción existente hasta el propio diseño del proceso. Es decir, se ha desarrollado la idea, iniciada por los expertos en el control de calidad japoneses, de que la contribución esencial de la Estadística debería consistir en determinar los modos efectivos de producción.

Por ejemplo, cuando se están fabricando chips de ordenadores, el productor necesita decidir sobre múltiples factores: las materias primas, la temperatura a la que se deben ensamblar las partes, la forma y el tamaño del chip, y otros factores. Para una elección determinada del conjunto de esos factores, el productor tendría que conocer la calidad media de los chips resultantes lo que le permitiría elegir los factores de producción que resultan más apropiados para obtener un producto de calidad.

En este capítulo se introduce la técnica estadística usada para analizar el anterior tipo de problemas. Consiste en un método general para hacer inferencias sobre una multitud de parámetros relacionados con las medias poblacionales. Su uso nos capacitará para determinar, por ejemplo, el nivel medio de calidad de un producto para las distintas elecciones posibles de los valores de los factores. La citada técnica estadística fue ideada por R.A. Fisher, y se conoce como el *análisis de la varianza* (ANOVA).

Mientras que el capítulo anterior abordaba los contrastes de las hipótesis que afectaban a las medias de dos poblaciones, en este capítulo se consideran los contrastes que afectan a múltiples medias poblacionales. Por ejemplo, en la sección 11.2 se supondrá que disponemos de datos procedentes de m poblaciones y que estamos interesados en contrastar la hipótesis de que todas las medias poblacionales son iguales. Este escenario se conoce como el análisis de la varianza unifactorial, puesto que el modelo asume que la media de una variable depende de un solo factor; esto es, depende de la muestra de la que se extrae la observación.

En la sección 11.3 se consideran modelos en los que se asume que dos factores determinan el valor medio de una variable. En tales casos, se puede concebir que las variables que se van a observar se pueden colocar en una tabla de doble entrada, y que el valor medio de una variable especificada depende tanto de la fila como de la columna en las que está situada. Para este problema de análisis de la varianza *bifactorial*, veremos cómo estimar los valores medios. Adicionalmente, mostraremos cómo contrastar la hipótesis de que un factor determinado no afecta a la media. Por ejemplo, podríamos disponer de datos sobre la cantidad de lluvia caída anualmente en varios puntos desérticos a lo largo de una serie de años. Dos factores podrían afectar a las cantidades anuales de lluvia en una región –la situación de la región y el año considerado– y podríamos estar interesados en contrastar si es sólo la situación, y no el año, lo que justifica las diferencias de lluvia anuales.

En todos los modelos considerados en este capítulo, se asumirá que los datos se distribuyen normalmente con la misma (aunque desconocida) varianza σ^2. El enfoque del análisis de la varianza para contrastar una hipótesis nula H_0 que afecta a múltiples parámetros se basa en obtener dos estimadores de la varianza común σ^2. El primero es un estimador válido de σ^2 tanto si la hipótesis nula es cierta como si no lo es, mientras que el segundo es un estimador válido sólo cuando H_0 es cierta. Adicionalmente, si H_0 no es cierta, este

último estimador sobreestima σ^2; esto es, tiende a sobrepasar la citada varianza. El contraste compara los valores de estos dos estimadores y rechaza H_0 cuando la razón del segundo estimador sobre el primero es suficientemente grande. En otras palabras, puesto que ambos estimadores deberían estar próximos cuando H_0 es cierta (ya que en este caso ambos estiman σ^2), mientras que el segundo estimador tiende a ser mayor que el primero cuando H_0 no es cierta, es natural rechazar H_0 cuando el segundo estimador es significativamente mayor que el primero.

11.2 Análisis de la varianza unifactorial

Consideremos m muestras, cada una de tamaño n. Supongamos que estas muestras son independientes y que la muestra i proviene de una población distribuida normalmente con una media μ_i y una varianza σ^2, $i = 1, \ldots, m$. Nos interesará contrastar la hipótesis nula

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_m$$

frente a

$$H_1: \text{no todas las medias son iguales}$$

Esto es, se tratará de contrastar la hipótesis nula de que todas las medias poblacionales son iguales frente a la alternativa de que al menos las medias de dos poblaciones difieren.

Denotemos por $\overline{X}_i$ y S_i^2 la media muestral y la varianza muestral, respectivamente, para los datos de la i-ésima muestra, $i = 1, \ldots, m$. Nuestro contraste de la hipótesis nula se llevará a cabo comparando los valores de dos estimadores de la varianza común σ^2. El primer estimador, que será un estimador válido de σ^2 tanto si la hipótesis nula es cierta como si no, se obtiene si se observa que cada una de las varianzas muestrales S_i^2 es un estimador insesgado de su varianza poblacional σ^2. Puesto que disponemos de m de estos estimadores, a saber, $S_1^2, \ldots, S_m^2$, se combinarán en un único estimador tomando el promedio de todos ellos. Esto es, el primer estimador de σ^2 viene dado por

$$\frac{1}{m}\sum_{i=1}^{m} S_i^2$$

Observe que este estimador se obtiene independientemente de que la hipótesis nula sea cierta o falsa.

El segundo estimador de σ^2 será un estimador válido sólo cuando la hipótesis nula sea cierta. Así pues, asumamos que H_0 es cierta, y que por tanto todas las medias poblacionales μ_i son iguales, digamos, $\mu_i = \mu$ para todo i. Bajo esta condición, se sigue que las m medias muestrales $\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_m$, estarán todas normalmente distribuidas con la misma media μ y la misma varianza σ^2/n. En otras palabras, cuando la hipótesis nula es cierta, los datos $\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_m$ constituyen una muestra procedente de una población normal con varianza σ^2/n. Denotemos esta varianza muestral por como $\overline{S}^2$. Esto es,

$$\overline{S}^2 = \frac{\sum_{i=1}^{m} (\overline{X}_i - \overline{\overline{X}})^2}{m - 1}$$

donde

$$\bar{\bar{X}} = \frac{1}{m}\sum_{i=1}^{m}\bar{X}_i$$

Puesto que $\bar{S}^2$ es un estimador insesgado de σ^2/n cuando H_0 es cierta, se sigue que en este caso $n\bar{S}^2$ es un estimador de σ^2. Esto es, el segundo estimador de σ^2 es $n\bar{S}^2$. De donde, se ha demostrado que

$$\sum_{i=1}^{m}\frac{S_i^2}{m} \qquad \text{siempre estima } \sigma^2$$

$$n\bar{S}^2 \qquad \text{estima } \sigma^2 \text{ cuando } H_0 \text{ es cierto}$$

Puesto que se puede demostrar que $n\bar{S}^2$ tiende a ser mayor que σ^2 cuando H_0 no es cierta, es razonable proponer que el estimador del contraste venga dado por

$$\text{TS} = \frac{n\bar{S}^2}{\sum_{i=1}^{m} S_i^2/m}$$

y rechazar H_0 cuando TS sea suficientemente grande.

Para determinar lo grande que TS debe ser para justificar el rechazo de H_0, se utiliza el hecho de que cuando H_0 es cierta, TS sigue la que se conoce como la *distribución F* con $m - 1$ grados de libertad en el numerador y $m(n - 1)$ grados de libertad en el denominador. Denotemos por $F_{m-1,\,m(n-1),\,\alpha}$ el α-valor crítico de esta distribución; es decir, la probabilidad de que una variable aleatoria F, con grados de libertad en el numerador y el denominador, respectivamente, iguales a $m - 1$ y $m(n - 1)$, sobrepase el valor $F_{m-1,\,m(n-1),\,\alpha}$ es igual a α (véase la figura 11.1). El contraste a nivel de significación α actuará como sigue:

$$\text{Rechazar } H_0 \qquad \text{si } \frac{n\bar{S}^2}{\sum_{i=1}^{m}\frac{S_i^2}{m}} \geq F_{m-1,\,m(n-1),\,\alpha}$$

$$\text{No rechazar } H_0 \qquad \text{en otro caso}$$

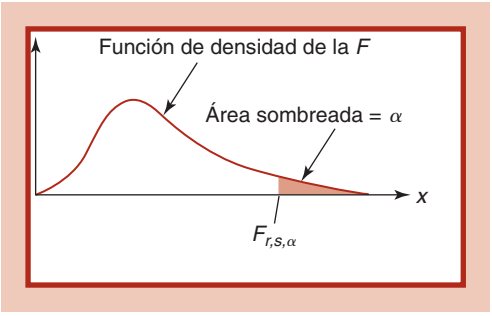


Figura 11.1 Variable aleatoria F con grados de libertad r, s: $P\{F \geq F_{r,s,a}\} = \alpha$.

Tabla 11.1 Valores de $F_{r,s,\,0,05}$

s = Grados de libertad del denominador	*r = Grados de libertad* del *numerador*			
	1	**2**	**3**	**4**
4	7,71	6,94	6,59	6,39
5	6,61	5,79	5,41	5,19
⋮	⋮	⋮	⋮	⋮
10	4,96	4,10	3,71	3,48

Los valores de $F_{r,\,s,\,0,05}$ para distintos valores de r y s, se incluyen el la tabla D.4 del Apéndice. Una parte de esta tabla se presenta en la tabla 11.1. Por ejemplo, en la tabla 11.1 se ve que existe una probabilidad del 5% de que una variable aleatoria F con 3 grados de libertad en el numerador y 10 en el denominador sobrepase 3,71.

Una observación sobre los grados de libertad

Los grados de libertad del numerador de una variable aleatoria F vienen determinados por el estimador del numerador $n\overline{S}^2$. Puesto que $\overline{S}^2$ es la varianza muestral de una muestra de tamaño m, sus grados de libertad son $m - 1$. Similarmente, el estimador del denominador se basa en el estadístico $\sum_{i=1}^{m} S_i^2$. Puesto que cada varianza muestral S_i^2 se calcula a partir de una muestra de tamaño n, cada una de ellas tiene $n - 1$ grados de libertad. La suma de las m varianzas muestrales conduce a un estadístico con $m\,(n - 1)$ grados de libertad.

Ejemplo 11.1 Un investigador de una cooperativa de consumidores diseñó un estudio de distancias asociadas a tres marcas distintas de gasolinas. Usó 15 motores idénticos ajustados para conseguir la misma velocidad, el investigador asignó aleatoriamente cada marca de gasolina a 5 de los citados motores. Después, se puso en marcha cada motor con 10 galones de gasolina. Las millas recorridas en cada caso fueron las siguientes.

Gasolina 1	Gasolina 2	Gasolina 3
220	244	252
251	235	272
226	232	250
246	242	238
260	225	256

Contraste la hipótesis de que la media de millas recorridas es igual para los tres tipos de gasolina. Utilice un nivel de significación del 5%.

Solución Puesto que hay tres muestras cada una de tamaño 5, se ve que $m = 3$ y $n = 5$. Las medias muestrales son

$$\bar{X}_1 = \frac{1203}{5} = 240{,}6$$
$$\bar{X}_2 = \frac{1178}{5} = 235{,}6$$
$$\bar{X}_3 = \frac{1268}{5} = 253{,}6$$

El promedio de las tres medias muestrales es

$$\bar{\bar{X}} = \frac{240{,}6 + 235{,}6 + 253{,}6}{3} = 243{,}2667$$

Por consiguiente, la varianza muestral de los datos $\bar{X}_i$, $i = 1, 2, 3$, es

$$\bar{S}^2 = \frac{(240{,}6 - 243{,}2667)^2 + (235{,}6 - 243{,}2667)^2 + (253{,}6 - 243{,}2667)^2}{2}$$
$$= 86{,}3333$$

El estimador del numerador es, pues,

$$5\bar{S}^2 = 431{,}667$$

Si se computan las varianzas muestrales de las tres muestras se obtiene $S_1^2 = 287{,}8$, $S_2^2 = 59{,}3$ y $S_3^2 = 150{,}8$, de donde el estimador del denominador es

$$\sum_{i=1}^{3} \frac{S_i^2}{3} = 165{,}967$$

Por consiguiente, el valor del estadístico del contraste es

$$\text{TS} = \frac{431{,}667}{165{,}967} = 2{,}60$$

Puesto que $m - 1 = 2$ y $m(n - 1) = 12$, debemos comparar el valor de TS con el valor de $F_{2,\,12,\,0{,}5}$. Ahora bien, a partir de la tabla D.4 del Apéndice, se ve que $F_{2,\,12,\,0{,}5} = 3{,}89$. Puesto que el valor del estadístico del contraste no sobrepasa 3,89, se sigue que, al nivel de significación del 5%, no se puede rechazar la hipótesis nula de que las gasolinas permiten recorrer iguales distancias en millas.

Otra forma de actuar para contrastar la hipótesis de que todas las medias poblacionales son iguales consiste en computar el p valor. Si el valor del estadístico del contraste, TS, es ν, el p valor vendrá dado por

$$p \text{ valor} = P\{F_{m-1,\,m(n-1)} \geq \nu\}$$

donde $F_{m-1,\, m(n-1)}$ representa una variable aleatoria F con $m - 1$ grados de libertad en el numerador y $m(n - 1)$ grados de libertad en el denominador. ■

El programa 11-1 permite computar el valor del estadístico del contraste, TS, y el correspondiente p valor.

Ejemplo 11.2 Hagamos los cálculos del ejemplo 11.1 usando el programa 11-1. Tras introducir los datos, se obtiene los siguientes resultados:

El estimador del denominador es 165,967
El estimador del numerador es 431,667
El valor del f-estadístico es 2,6009
El p valor es 0,11525 ■

La tabla 11.2 sintetiza los resultados de esta sección.

Observación: Cuando $m = 2$, lo anterior equivale a un contraste cuya hipótesis nula es que dos muestras independientes, con varianzas poblacionales iguales, provienen de poblaciones con medias idénticas. El lector podría preguntarse cómo, en este caso, se puede comparar el procedimiento indicado con el presentado en el capítulo 10. Resulta que los dos contrastes son exactamente idénticos. Esto es, si se usan los mismos datos ambos procedimientos conducen exactamente al mismo p valor.

Tabla 11.2 Tabla ANOVA unifactorial

Las variables $\overline{X}_i$ *y* $S_i^2, i = 1, \ldots, m$, *son, respectivamente, las medias muestrales y las varianzas muestrales de muestras independientes de tamaño n procedentes de poblaciones normales con medias* μ_i *y varianzas comunes* σ^2.

Fuente del estimador	Estimador de σ^2	Valor del estadístico del contraste
Entre muestras	$n\overline{S}^2 = \dfrac{n\sum_{i=1}^{m}(\overline{X}_i - \overline{\overline{X}})^2}{m-1}$	$\text{TS} = \dfrac{n\overline{S}^2}{\sum_{i=1}^{m}\frac{S_i^2}{m}}$
Dentro de las muestras	$\sum_{i=1}^{m}\dfrac{S_i^2}{m}$	

Contraste, a nivel de significación α, de H_0: todos los valores μ_i son iguales:

Rechazar H_0	si $\text{TS} \geq F_{m-1,\, m(n-1),\, \alpha}$
No rechazar H_0	en otro caso

Si $\text{TS} = \nu$, entonces p valor $= P\{F_{m-1,\, m(n-1)} \geq \nu\}$

donde $F_{m-1,\, m(n-1)}$ representa una variable aleatoria F con $m - 1$ grados de libertad en el numerador y $m(n - 1)$ grados de libertad en el denominador.

Problemas

1. Consideremos los datos de tres muestras, cada una de tamaño 4. (Esto es, $m = 3, n = 4$.)

Muestra 1	5	9	12	6
Muestra 2	13	12	20	11
Muestra 3	8	12	16	8

   (a) Calcule las tres medias muestrales $\overline{X}_i$, $i = 1, 2, 3$.

   (b) Encuentre $\overline{\overline{X}}$, la media de las tres medias muestrales.

   (c) Compruebe que $\overline{\overline{X}}$ es igual a la media de los 12 valores de datos.

2. Utilice los datos del problema 1 para contrastar la hipótesis de que las medias de las tres poblaciones son iguales. Tome un nivel de significación del 5%.

3. Un nutricionista dividió aleatoriamente a 15 ciclistas en tres grupos de 5 unidades cada uno. Los miembros del primer grupo recibieron suplementos vitamínicos, que añadieron a sus comidas durante las tres semanas siguientes. Al los del segundo grupo se les indicó que, durante esas tres semanas, tomaran un tipo especial de cereal de grano entero rico en fibra. A los miembros del tercer grupo se les dijo que comieran como hacían normalmente. Transcurrido el periodo indicado el nutricionista hizo que cada ciclista recorriera una distancia de 6 millas. Se registraron los siguientes tiempos:

Grupo de las vitaminas	15,6	16,4	17,2	15,5	16,3
Grupo del cereal rico en fibra	17,1	16,3	15,8	16,4	16,0
Grupo de control	15,9	17,2	16,4	15,4	16,8

   ¿Son estos datos consistentes con la hipótesis de que ni las vitaminas ni el cereal rico en fibra afectan a la velocidad de los ciclistas? Use un nivel de significación del 5%.

4. Para determinar si el porcentaje de calorías por consumo de grasas en la dieta de una persona es igual en todo el país, se seleccionaron muestras aleatorias de 20 voluntarios en tres regiones diferentes. Se determinó el porcentaje de calorías procedentes de grasas en cada voluntario, resultando los siguientes datos sumariales:

Región *i*	$\overline{X}_i$	S_i^2
$i = 1$	32,4	102
$i = 2$	36,4	112
$i = 3$	37,1	138

   Contraste la hipótesis de que el porcentaje de calorías procedentes de grasas no varía en los individuos que viven en esas tres regiones. Use un nivel de significación del 5%.

5. Se analizaron los contenidos en grasa de seis raciones de tres marcas diferentes de carne procesada. Resultaron los siguientes datos (en porcentaje de grasa por gramo de peso):

Marca	Contenido de grasa					
1	32	34	31	35	33	30
2	40	36	33	29	35	32
3	37	30	28	33	37	39

Al nivel de significación del 5%, ¿permiten estos datos rechazar la hipótesis de que los contenidos de grasa medios son los mismos para las tres marcas?

6. Un factor importante en las ventas de una nueva bola de golf es la distancia puede alcanzar tras ser golpeada. Con una máquina automática de golpeo se dieron 25 golpes con cuatro tipos distintos de bolas, y se registraron las distancias conseguidas (en yardas). Resultaron los siguientes datos, referentes a las medias muestrales y a las varianzas muestrales.

Tipo de bola i	$\overline{X}_i$	S_i^2
1	212	26
2	220	23
3	198	25
4	214	24

Al nivel de significación del 5%, contraste la hipótesis nula de que la distancia media conseguida es la misma para todos los tipos de bola.

7. Se utilizaron tres procedimientos químicos estándar para determinar el contenido de magnesio de un cierto compuesto químico. Cada procedimiento se aplicó 4 veces sobre un compuesto dado, y resultaron los datos siguientes:

Método 1	76,43	78,61	80,40	78,22
Método 2	80,40	82,24	72,70	76,04
Método 3	82,16	84,14	80,20	81,33

Contraste la hipótesis de que las lecturas medias son iguales con los tres métodos. Use un nivel de significación del 5%.

8. Una médica de urgencias desea saber si existen diferencias entre los tiempos que tres diferentes esteroides por inhalación tardan en curar un ataque asmático benigno. Durante varias semanas, la médica administró aleatoriamente estos tres esteroides a pacientes con asma y anotó el número de minutos que tardaron en aclarase los pulmones de los pacientes. Cada tipo de esteroide se aplicó a 12 pacientes, y se obtuvieron las siguientes medias muestrales y varianzas muestrales:

Esteroide	$\overline{X}_i$	S_i^2
A	32	145
B	40	138
C	30	150

Contraste la hipótesis de que el tiempo medio que se tarda en sofocar un ataque asmático benigno es el mismo con los tres esteroides. Use un nivel de significación del 5%.

9. Los siguientes datos se refieren al número de muertos por cada 10 000 adultos en una gran ciudad durante las distintas estaciones de los años entre 1982 y 1986.

Año	Invierno	Primavera	Verano	Otoño
1982	33,6	31,4	29,8	32,1
1983	32,5	30,1	28,5	29,9
1984	35,3	33,2	29,5	28,7
1985	34,4	28,6	33,9	30,1
1986	37,3	34,1	28,5	29,4

Contraste la hipótesis de que las tasas de mortalidad no dependen de la estación del año. Utilice un nivel de significación del 5%.

10. Un experto en nutrición piensa que las distancias que recorren normalmente los corredores amateurs no están relacionadas con sus niveles de colesterol en la sangre. Se eligió aleatoriamente a seis corredores de tres categorías distintas (en número de millas corridas semanalmente) y se chequearon sus niveles de colesterol en la sangre. Las medias muestrales y las varianzas muestrales que se obtuvieron son las siguientes

Millas semanales recorridas	$\overline{X}_i$	S_i^2
Menos de 15	188	190
Entre 15 y 30	181	211
Más de 30	174	202

¿Prueban estos datos la idea del nutricionista? Utilice un nivel de significación del 5%.

11. Un administrador universitario mantiene que las calificaciones medias del primer curso universitario no presentan diferencias entre los estudiantes de los tres institutos locales. Los datos siguientes muestran las calificaciones medias obtenidas por 15 estudiantes en el primer año universitario eligidos aleatoriamente, 5 de cada uno de los institutos locales.¿Son estos datos lo suficientemente fuertes para contradecir, al nivel del 5%, la idea del administrador?

Instituto A	Instituto B	Instituto C
3,2	2,8	2,5
2,7	3,0	2,8
3,0	3,3	2,4
3,3	2,5	2,2
2,6	3,1	3,0

12. Un psicólogo diseñó un experimento a partir del cual se obtenían las puntuaciones de un test de laberinto al que fueron sometidos unos ratones que habían sido entrenados bajo distintas condiciones de laboratorio. Un conjunto de 24 se dividió aleatoriamente en tres grupos de 8 cada uno. A los miembros del primer grupo se les sometió un tipo de entrenamiento cognitivo, a los del segundo grupo se les sometió a un cierto tipo de entrenamiento de conducta, y los miembros del tercer grupo no fueron entrenados en absoluto. Las puntuaciones del test de laberinto (asignadas por alguien que desconocía el entrenamiento al que había sido sometido cada ratón) fueron las siguientes.

Grupo	$\overline{X}_i$	S_i^2
1	74,2	111,4
2	78,5	102,1
3	80,0	124,0

¿Existe suficiente evidencia como para concluir que los distintos tipos de entrenamiento afectan a las puntuaciones del test de laberinto? Use un nivel de significación del 5%.

11.3 Análisis de la varianza bifactorial: introducción y estimación de parámetros

El modelo de la sección 11.2 nos permite estudiar el efecto de un solo factor sobre un conjunto de datos, pero también se pueden estudiar los efectos de varios factores. En esta sección se supondrá que existen dos factores que afectan a cada dato.

Ejemplo 11.3 Cinco estudiantes se sometieron a cuatro tests de lectura diferentes. Sus puntuaciones fueron las siguientes:

	Estudiante				
Examen	**1**	**2**	**3**	**4**	**5**
1	75	73	60	70	86
2	78	71	64	72	90
3	80	69	62	70	85
4	73	67	63	80	92

En este conjunto de 20 datos, hay dos factores que afectan a cada valor: el examen y el estudiante cuya puntuación se registró en dicho examen. El factor examen tiene cuatro valores, o *niveles*, posibles, mientras que el factor estudiante presenta cinco niveles posibles. ■

En general, supongamos que existen m valores posibles del primer factor y n valores posibles del segundo. Denotemos por X_{ij} el valor del dato obtenido cuando el primer factor

está en el nivel i y el segundo está en el nivel j. Se puede representar el conjunto de datos mediante la siguiente tabla de doble entrada (por filas y columnas):

$$\begin{array}{ccccccc} X_{11} & X_{12} & \cdots & X_{1j} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2j} & \cdots & X_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ X_{i1} & X_{i2} & \cdots & X_{ij} & \cdots & X_{in} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ X_{m1} & X_{m2} & \cdots & X_{mj} & \cdots & X_{mn} \end{array}$$

Por este motivo, el primer factor se identificará como factor *fila*, y el segundo como factor *columna*. De esta forma, X_{ij} es el valor correspondiente a la fila i y a la columna j.

Como en la sección 11.2, se supondrá que todos los valores X_{ij}, $i = 1, \ldots, m, j = 1, \ldots, n$, son variables aleatorias normales e independientes con una varianza común σ^2. Sin embargo, mientras que en la sección 11.2 se supuso que sólo un factor afectaba al valor medio de un dato –a saber, la muestra a la cual pertenecía– en esta sección se asumirá que el valor medio de un dato depende tanto de su fila como de su columna. En cualquier caso, antes de especificar este modelo, debemos reconsiderar el modelo de la sección 11.2. Si hacemos que X_{ij} represente el valor del j-ésimo miembro de la muestra i, este último modelo asume que

$$E[X_{ij}] = \mu_i$$

Si se denota con una μ la media de los valores μ_i, es decir,

$$\mu = \frac{\sum_{i=1}^{m} \mu_i}{m}$$

lo anterior se puede escribir como

$$E[X_{ij}] = \mu + \alpha_i$$

donde $\alpha_i = \mu_i - \mu$. Con esta definición α_i como la desviación de μ_i del promedio de las medias, μ, es fácil ver que

$$\sum_{i=1}^{m} \alpha_i = 0$$

En el caso de dos factores, escribiremos nuestro modelo en términos de las desviaciones por filas y por columnas. Específicamente, se supondrá que el valor esperado de la variable X_{ij} se puede expresar como sigue:

$$E[X_{ij}] = \mu + \alpha_i + \beta_j$$

El valor μ se denomina *media total*, α_i es la *desviación de la media total debida a la fila i*, y β_j es la *desviación de la media total debida a la columna j*.

Adicionalmente, estas desviaciones cumplen las igualdades siguientes.

$$\sum_{i=1}^{m}\alpha_i = \sum_{j-1}^{n}\beta_j = 0$$

Empecemos determinando los estimadores de los parámetros, μ, α_i y β_j, $i = 1, \ldots, m$, $j = 1, \ldots, n$. Para hacerlo, será conveniente introducir la siguiente notación "con puntos". Sea

$$X_{i\cdot} = \frac{\sum_{j=1}^{n} X_{ij}}{n} = \text{media de todos los valores de la fila } i$$

$$X_{\cdot j} = \frac{\sum_{i=1}^{m} X_{ij}}{m} = \text{media de todos los valores de la columna } j$$

$$X_{\cdot\cdot} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}}{nm} = \text{media de todos los } nm \text{ valores}$$

No es difícil comprobar que

$$E[X_{i\cdot}] = \mu + \alpha_i$$
$$E[X_{\cdot j}] = \mu + \beta_j$$
$$E[X_{\cdot\cdot}] = \mu$$

Puesto que lo anterior es equivalente a

$$E[X_{\cdot\cdot}] = \mu$$
$$E[X_{i\cdot} - X_{\cdot\cdot}] = \alpha_i$$
$$E[X_{\cdot j} - X_{\cdot\cdot}] = \beta_j$$

se ve que los estimadores insesgados de μ, α_i y β_j—llamémosles $\hat{\mu}$, $\hat{\alpha}_i$ y $\hat{\beta}_j$—vienen dados por

$$\hat{\mu} = X_{\cdot\cdot}$$

$$\hat{\alpha}_i = X_{i\cdot} - X_{\cdot\cdot}$$

$$\hat{\beta}_j = X_{\cdot j} - X_{\cdot\cdot}$$

Ejemplo 11.4 Los siguientes datos, del ejemplo 11.3, muestran las puntuaciones obtenidas por cinco estudiantes que fueron sometidos a cuatro tests de lectura diferentes. Utilícelos para estimar los parámetros del modelo.

	Estudiante						
Exámen	**1**	**2**	**3**	**4**	**5**	**Totales por fila**	$X_{i\cdot}$
1	75	73	60	70	86	364	72,8
2	78	71	64	72	90	375	75
3	80	69	62	70	85	366	73,2
4	73	67	63	80	92	375	75
Totales por columna	306	280	249	292	353	1480	← total general
$X_{\cdot j}$	76,5	70	62,25	73	88,25	$X_{\cdot\cdot} = \frac{1480}{20} = 74$	

Los estimadores son

$$\hat{\mu} = 74$$

$$\hat{\alpha}_1 = 7{,}8 - 74 = -1{,}2 \qquad \hat{\beta}_1 = 76{,}5 - 74 = 2{,}5$$

$$\hat{\alpha}_2 = 75 - 74 = 1 \qquad \hat{\beta}_2 = 70 - 74 = -4$$

$$\hat{\alpha}_3 = 73{,}2 - 74 = -0{,}8 \qquad \hat{\beta}_3 = 62{,}25 - 74 = -11{,}75$$

$$\hat{\alpha}_4 = 75 - 74 = 1 \qquad \hat{\beta}_4 = 73 - 74 = -1$$

$$\hat{\beta}_5 = 88{,}25 - 74 = 14{,}25$$

En consecuencia, si se elige a uno de los estudiantes aleatoriamente y después se le somete a un examen elegido también aleatoriamente, nuestro estimador de la puntuación media que se obtendría es $\hat{\mu} = 74$. Si se nos dijera que el estudiante fue sometido al examen i, ello incrementaría nuestro estimador de la puntuación media en la cantidad $\hat{\alpha}_i$; y si se nos dijera que el estudiante elegido fue el número j, ello incrementaría nuestro estimador de la puntuación media la cantidad $\hat{\beta}_j$. Así pues, estimaríamos, por ejemplo, que la puntuación obtenida en el examen 1 por el estudiante 2 es el valor de una variable aleatoria cuya media es $\hat{\mu} + \hat{\alpha}_1 + \hat{\beta}_2 = 74 - 1{,}2 - 4 = 68{,}8$. ■

Observación: En lo anterior se ha definido $X_{\cdot\cdot}$ usando una notación de doble sumatorio. Esto es, se ha utilizado la notación de la forma

$$\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}$$

Esta expresión significa que se debe sumar los términos X_{ij} para todos los valores posibles de los pares i,j.

Equivalentemente, supongamos que los valores de los datos X_{ij} se presentan en una tabla de doble entrada (por filas y columnas) tal como se hizo al comienzo de esta sección. Denotemos por T_i la suma de los valores de la fila i. Esto es,

$$T_i = \sum_{j=1}^{n} X_{ij}$$

De acuerdo con esto, la notación de doble sumatorio viene dada por

$$\sum_{i=1}^{m}\sum_{j=1}^{n}X_{ij} = \sum_{i=1}^{m}T_i$$

Dicho de otro modo, el doble sumatorio es igual a la suma de todas las sumas por filas; esto es, coincide con la suma de todos los nm valores X_{ij}. (Es fácil ver que también es igual a la suma de todas las sumas por columnas.)

Problemas

1. Para un estudio sobre la polución del aire, se tomaron muestras de aire en tres lugares diferentes y en cinco fechas distintas. Los datos siguientes se refieren a la cantidad de partículas en suspensión presentes en el aire (en unidades de miligramos por metro cúbico).

	Lugares		
Fechas	**1**	**2**	**3**
1. Enero de 2001	78	84	87
2. Julio de 2001	75	69	82
3. Enero de 2002	66	60	70
4. Julio de 2002	71	64	61
5. Enero de 2003	58	55	52

Suponiendo el modelo

$$E[X_{ij}] = \mu + \alpha_i + \beta_j$$

estime los parámetros desconocidos.

2. Con los datos del problema 1, verifique que

$$X_{..} = \frac{\sum_{i=1}^{m}X_{i\cdot}}{m} = \frac{\sum_{j=1}^{n}X_{j\cdot}}{n}$$

Exlique qué establece esta ecuación.

3. Los siguientes datos indican el número de cajas que tres hombres han empaquetado individualmente en tres turnos de trabajo distintos.

	Hombre		
Turno	**1**	**2**	**3**
1. De 9 a 11 h.	32	27	29
2. De 13 a 15 h.	31	26	22
3. De 15 a 17 h.	33	30	25

Con el modelo presentado en esta sección, estime los parámetros desconocidos.

4. Utilice los resultados del ejemplo 11.4 para estimar $E[X_{ij}] = \mu + \alpha_i + \beta_j$ para todos los posibles valores de i y j, $i = 1, 2, 3, 4, j = 1, 2, 3, 4, 5$. Compare los valores de $E[X_{ij}]$ con los valores observados de X_{ij} dados en dicho ejemplo.

5. La tabla siguiente presenta las tasas de natalidad por cada 1000 habitantes de cuatro países diferentes en cuatro años distintos.

	2003	**2002**	**2001**	**1990**
Australia	12,6	12,71	12,86	15,4
Austria	9,4	9,58	9,74	11,6
Bélgica	10,4	10,58	10,74	12,6
República Checa	9,0	9,08	9,11	13,4

Asumiendo el modelo de esta sección, estime:

(a) La media total de las tasas de natalidad

(b) La desviación de la media total de las tasas de natalidad australianas.

(c) La desviación de la media total de las tasas de natalidad correspondientes a 1990.

6. En la tabla siguiente se muestran las tasas de desempleo en tres niveles educacionales en cuatro años distintos.

Nivel de educación	**1980**	**1984**	**1988**	**2000**
Sin educación secundaria	8,4	12,1	9,6	8,8
Con educación secundaria	5,1	7,2	5,4	6,1
Graduado universitario	1,9	2,7	1,7	2,2

Fuente: Oficina de Estadísticas de Empleo de Estados Unidos, *Estadísticas de Empleo.*

Con el modelo de esta sección, estime:

(a) La media total μ

(b) Las desviaciones por filas, α_i, $i = 1, 2, 3, 4$

(c) Las desviaciones por columnas, β_j, $j = 1, 2, 3, 4$

7. La tabla siguiente muestra las tasas de desempleo de cinco actividades industriales en tres años distintos.

Actividad industrial	2000	2001	2002
Transporte	3,4	4,3	4,9
Minería	4,4	4,2	6,3
Construcción	6,2	7,1	9,2
Manufacturación	3,5	5,2	6,7
Información	3,2	4,9	6,9

Fuente: Oficina de Estadísticas de Empleo de Estados Unidos, *Empleo y salarios.*.

Con el modelo de esta sección, estime los parámetros desconocidos.

8. Suponga que $x_{ij} = i + 4j$. (Así, por ejemplo, $x_{11} = 1 + 4 = 5$ y $x_{23} = 2 + 4 \cdot 3 = 14$.) Escriba en una tabla de doble entrada los 12 valores de x_{ij}, donde i es 1 o 2 o 3 y j es 1 o 2 o 3 o 4, poniendo x_{ij} en la celda correspondiente a la fila i y a la columna j.

9. En el problema 8, determine:

(a) $\sum_{j=1}^{4} x_{1j}$ (b) $\sum_{j=1}^{4} x_{2j}$

(c) $\sum_{j=1}^{4} x_{3j}$ (d) $\sum_{i=1}^{3}\sum_{j=1}^{4} x_{ij}$

11.4 Análisis de la varianza bifactorial: contraste de hipótesis

Consideremos el modelo bifactorial en el que se dispone de los datos X_{ij}, $i = 1, \ldots, m$ y $j = 1, \ldots, n$. Se asumirá que estos datos son variables aleatorias normales e independientes con varianza común σ^2 y con valores medios que cumplen

$$E[X_{ij}] = \mu + \alpha_i + \beta_j$$

donde

$$\sum_{i=1}^{m} \alpha_i = \sum_{j=1}^{n} \beta_j = 0$$

En esta sección se contrastarán las hipótesis

$$H_0: \alpha_i = 0 \text{ para todo } i$$

frente a

$$H_1: \text{no todos los } \alpha_i \text{ son } 0$$

La hipótesis nula establece que no existe el efecto fila, en el sentido de que el valor de un dato no se ve afectado por su nivel del factor fila.

También se contrastarán las hipótesis análogas por columnas, es decir,

$$H_0: \beta_j = 0 \text{ para todo } j$$

frente a

$$H_1: \text{no todos los } \beta_j \text{ son } 0$$

Para contrastar las hipótesis precedentes, se llevará a cabo el enfoque propio del análisis de la varianza, según el cual se obtendrán dos estimadores distintos de la varianza σ^2. El primero siempre será un estimador válido, mientras que el segundo sólo será un estimador válido cuando la hipótesis nula sea cierta. Adicionalmente, el segundo estimador tenderá a sobreestimar σ^2 cuando la hipótesis nula no sea cierta.

Para obtener nuestro primer estimador de σ^2, recordemos que la suma de cuadrados de N variables aleatorias independientes normales estándar es una variable aleatoria chi-cuadrado con N grados de libertad. Puesto que las nm variables estandarizadas

$$\frac{X_{ij} - E[X_{ij}]}{\sigma}$$

$i = 1, \ldots, m, j = 1, \ldots, n$ son todas normales estándar e independientes, se sigue que

$$\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} - E[X_{ij}])^2}{\sigma^2} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} - \mu - \alpha_i - \beta_j)^2}{\sigma^2}$$

es una chi-cuadrado con nm grados de libertad. Si en la expresión anterior reemplazamos ahora los parámetros desconocidos $\mu, \alpha_1, \alpha_2, \ldots, \alpha_m, \beta_1, \beta_2, \ldots, \beta_n$ por sus estimadores $\hat{\mu}, \hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_m, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_n$, la expresión resultante continúa siendo una chi-cuadrado aunque perderá un grado de libertad por cada parámetro estimado. Para determinar cuántos parámetros se han de estimar debemos ser cuidadosos y recordar que $\sum_{i=1}^{m}\alpha_i = \sum_{j=1}^{n}\beta_j = 0$. Puesto que la suma de todos los α_i es 0,se sigue que una vez que hayamos estimado $m - 1$ de los α_i, también habremos estimado el último. Por consiguiente, solamente hay que estimar $m - 1$ parámetros α_i para determinar todos los $\hat{\alpha}_i$. Por la misma razón, sólo es necesario estimar $n - 1$ de los β_j para tener estimados los n. Puesto que μ también se debe estimar, se ve que el número total de parámetros a estimar es

$$1 + (m - 1) + (n - 1) = m + n - 1$$

Como resultado, se sigue que

$$\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j)^2}{\sigma^2}$$

es una variable aleatoria chi-cuadrado con $nm - (n + m - 1) = (n - 1)(m - 1)$ grados de libertad.

Puesto que

$$\hat{\mu} = X..$$

$$\hat{\alpha}_i = X_{i\cdot} - X..$$

$$\hat{\beta}_j = X_{\cdot j} - X..$$

se ve que

$$\begin{aligned}\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j &= X.. + X_{i\cdot} - X.. + X_{\cdot j} - X.. \\ &= X_{i\cdot} + X_{\cdot j} - X..\end{aligned}$$

Por tanto, el estadístico

$$\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} - X_{i\cdot} - X_{\cdot j} + X..)^2}{\sigma^2} \tag{11.1}$$

es una chi-cuadrado con $(n - 1)(m - 1)$ grados de libertad.

La suma de cuadrados SS_e definida por

$$SS_e = \sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} - X_{i\cdot} - X_{\cdot j} + X..)^2$$

se conoce como *suma de cuadrados de los errores*.

Si la diferencia entre una variable aleatoria y su media estimada se concibe como un "error", entonces SS_e es igual a la suma de cuadrados de los errores. Puesto que SS_e/σ^2 es precisamente la expresión de la ecuación (11.1), se ve que SS_e/σ^2 es una chi-cuadrado con $(n - 1)(m - 1)$ grados de libertad. Como el valor esperado de una variable aleatoria chi-cuadrado es igual al número de sus grados de libertad, se tiene que

$$E\left[\frac{\mathrm{SS}_e}{\sigma^2}\right] = (n - 1)(m - 1)$$

o

$$E\left[\frac{SS_e}{(n - 1)(m - 1)}\right] = \sigma^2$$

Esto es, haciendo $N = (n - 1)(m - 1)$, se ha probado lo siguiente:

$\frac{SS_e}{N}$ es un estimador insesgado de σ^2.

Supongamos ahora que se quiere contrastar la hipótesis nula de que no existe efecto fila; es decir, se quiere contrastar

$$H_0: \alpha_i = 0 \text{ para todo } i$$

frente a

$$H_1: \text{no todos los } \alpha_i \text{ son } 0$$

Para obtener un segundo estimador de σ^2, consideremos las medias por filas $X_{i\cdot}$, $i = 1, \ldots, m$. Observe que cuando H_0 es cierta, cada α_i es igual a 0, y, por tanto,

$$E[X_{i\cdot}] = \mu + \alpha_i = \mu$$

Puesto que cada $X_{i\cdot}$ es la media de n variables aleatorias, cada una de las cuales tiene una varianza σ^2, se sigue que

$$\text{Var}(X_{i\cdot}) = \frac{\sigma^2}{n}$$

Así pues, se ve que cuando H_0 es cierta,

$$\frac{\sum_{i=1}^{m}(X_{i\cdot} - E[X_{i\cdot}])^2}{\text{Var}(X_{i\cdot})} = \frac{n\sum_{i=1}^{m}(X_{i\cdot} - \mu)^2}{\sigma^2}$$

seguirá una chi-cuadrado con m grados de libertad. Si, ahora, en lo anterior se sustituye μ por $X_{\cdot\cdot}$ (el estimador de μ) la expresión resultante continuará siendo una chi-cuadrado aunque con un grado de libertad menos. Esto es, tendrá $m - 1$ grados de libertad. Así pues, se ha obtenido lo siguiente.

Cuando H_0 es cierta,

$$\frac{n\sum_{i=1}^{m}(X_{i\cdot} - X_{\cdot\cdot})^2}{\sigma^2}$$

es una chi-cuadrado con $m - 1$ grados de libertad

Al estadístico SS_e, definido por

$$\mathrm{SS}_r = n\sum_{i=1}^{m}(X_{i\cdot} - X_{\cdot\cdot})^2$$

se le conoce como *suma de cuadrados de las filas*.

Se ha visto, pues, que cuando H_0 es cierta, SS_r/σ^2 es una chi-cuadrado con $m - 1$ grados de libertad. Como resultado, cuando H_0 es cierta,

$$E\left[\frac{\mathrm{SS}_r}{\sigma^2}\right] = m - 1$$

o, equivalentemente,

$$E\left[\frac{\mathrm{SS}_r}{m-1}\right] = \sigma^2$$

Adicionalmente, se puede demostrar que $\mathrm{SS}_r/(m-1)$ tenderá a ser mayor que σ^2 cuando H_0 no es cierta. Así pues, de nuevo se han obtenido dos estimadores de σ^2. El primer estimador, SS_e/N, donde $N = (n-1)(m-1)$, es un estimador válido tanto si la hipótesis nula es cierta como si no. El segundo estimador, $\mathrm{SS}_r/(m-1)$,es un estimador válido de σ^2 sólo cuando H_0 es cierta y tiende a ser mayor que σ^2 cuando H_0 no es cierta.

El contraste de la hipótesis nula H_0 de que no existe efecto fila implica comparar los dos estimadores dados y rechazar H_0 cuando el segundo es significativamente mayor que el primero. Específicamente, se utiliza el estadístico del contraste

$$\mathrm{TS} = \frac{\mathrm{SS}_r/(m-1)}{\mathrm{SS}_e/N}$$

y el contraste a nivel de significación α consiste en

Rechazar H_0	si $\mathrm{TS} \geq F_{m-1,\,N,\,\alpha}$
No rechazar H_0	en otro caso

Alternativamente, el contraste se puede realizar si se calcula el p valor. Si el valor del estadístico del contraste es v, el p valor viene dado por

$$p \text{ valor} = P\{F_{m-1,\,N} \geq v\}$$

donde $F_{m-1,N}$ es una variable aleatoria F con $m - 1$ grados de libertad en el numerador y N grados de libertad en el denominador.

De igual forma, se puede contrastar la hipótesis nula de que no existe el efecto columna; esto es, de que todos los β_j son iguales a 0. Los resultados de ambos contrastes aparecen sintetizados en la tabla 11.3.

El programa 11-2 realiza todos los cálculos y obtiene el p valor.

Tabla 11.3 ANOVA bifactorial

	Suma de cuadrados	Grados de libertad
De las filas	$\text{SS}_r = n\sum_{i=1}^{m}(X_{i\cdot} - X_{\cdot\cdot})^2$	$m - 1$
De las columna	$\text{SS}_c = m\sum_{j=1}^{n}(X_{\cdot j} - X_{\cdot\cdot})^2$	$n - 1$
Error	$\text{SS}_e = \sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} - X_{i\cdot} - X_{\cdot j} + X_{\cdot\cdot})^2$	$(n - 1)(m - 1)$
	Siendo $N = (n - 1)(m - 1)$	

Hipótesis nula	Estadístico del contraste	Contraste a nivel de significación α	p valor si TS = v
No existen efectos fila ($\alpha_i = 0$ para todo i)	$\dfrac{SS_r/(m-1)}{\text{SS}_e/N}$	Rechazar si TS $\geq F_{m-1,\,N,\,\alpha}$	$P\{F_{m-1,\,N} \geq v\}$
No existen efectos columna ($\beta_j = 0$ para todo j)	$\dfrac{SS_c/(n-1)}{\text{SS}_e/N}$	Rechazar si TS $\geq F_{n-1,\,N,\,\alpha}$	$P\{F_{n-1,\,N} \geq v\}$

Ejemplo 11.5 Los siguientes datos representan el número de piezas defectuosas producidas por cuatro trabajadores que utilizan, por turnos, tres máquinas diferentes.

	Trabajador			
Máquina	**1**	**2**	**3**	**4**
1	41	42	40	35
2	35	42	43	36
3	42	39	44	47

Contraste si existen diferencias significativas entre las máquinas y los trabajadores.

Solución Puesto que existen tres filas y cuatro columnas, se ve que $m = 3$ y $n = 4$. Si se calculan las medias por filas y por columnas se obtienen los siguientes resultados:

$$X_{1\cdot} = \frac{41 + 42 + 40 + 35}{4} = 39{,}5 \qquad X_{\cdot 1} = \frac{41 + 35 + 42}{3} = 39{,}33$$

$$X_{2\cdot} = \frac{35 + 42 + 43 + 36}{4} = 39 \qquad X_{\cdot 2} = \frac{42 + 42 + 39}{3} = 41$$

$$X_{3\cdot} = \frac{42 + 39 + 44 + 47}{4} = 43 \qquad X_{\cdot 3} = \frac{40 + 43 + 44}{3} = 42{,}33$$

$$X_{\cdot 4} = \frac{35 + 36 + 47}{3} = 39{,}33$$

Adicionalmente,

$$X.. = \frac{39{,}5 + 39 + 43}{3} = 40{,}5$$

Por consiguiente,

$$\begin{aligned} \mathrm{SS}_r &= n\sum_{i=1}^{m}(X_{i\cdot} - X..)^2 \\ &= 4[1^2 + (1{,}5)^2 + (2{,}5)^2] \\ &= 38 \end{aligned}$$

y

$$\begin{aligned} \mathrm{SS}_c &= m\sum_{j=1}^{n}(X_{\cdot j} - X..)^2 \\ &= 3[(1{,}17)^2 + (0{,}5)^2 + (1{,}83)^2 + (1{,}17)^2] \\ &= 19{,}010 \end{aligned}$$

El cálculo de SS_e es más pesado porque se han de sumar los cuadrados de los términos $X_{ij} - X_{i\cdot} - X_{\cdot j} + X..$, cuando i varía del 1 al 3 y j del 1 al 4. El primer término de esta suma, cuando $i = 1$ y $j = 1$, es

$$(41 - 39{,}5 - 39{,}33 + 40{,}5)^2$$

Si sumamos los 12 términos se obtiene

$$\mathrm{SS}_e = 94{,}05$$

Puesto que $m - 1 = 2$ y $N = 2 \cdot 3 = 6$, para contrastar la hipótesis de que no existe efecto fila el estadístico del contraste será

$$\mathrm{TS(filas)} = \frac{38/2}{94{,}05/6} = 1{,}21$$

De la tabla D.4 del Apéndice, se ve que $F_{2,\,6,\,0{,}05} = 5{,}14$. Por tanto, la hipótesis de que la máquina que se usa no afecta al número medio de piezas defectuosas no resulta rechazada al nivel de significación del 5%.

Para contrastar la hipótesis de que no existe el efecto columna el estadístico del contraste es

$$\mathrm{TS(columnas)} = \frac{19{,}010/3}{94{,}05/6} = 0{,}40$$

De la tabla D.4 del Apéndice, se ve que $F_{3,\,6,\,0{,}05} = 4{,}76$, por consiguiente, tampoco se rechaza la hipótesis de que el número medio de piezas defectuosas no se ve afectado por el trabajador que la usa, al nivel de significación del 5%. ■

El ejemplo anterior también se podría resolver también utilizando el programa 11-2. Este programa produce la siguiente salida:

El valor del F-estadístico para contrastar que no existe efecto fila es 1,212766.

El *p* valor para contrastar que no existe efecto fila es 0,3571476.

El valor del F-estadístico para contrastar que no existe efecto columna es 0,4042554.

El *p* valor para contrastar que no existe efecto columna es 0,7555629.

Puesto que ambos *p* valores son superiores a 0,05, no se puede rechazar, al nivel de significación del 5%, la hipótesis de que la máquina usada no afecta al número de piezas defectuosas producidas; ni tampoco se puede rechazar la hipótesis de que el trabajador empleado no afecta al número medio de piezas defectuosas producidas.

Problemas

1. Se ha llevado a cabo un experimento para determinar el efecto de tres combustibles distintos y de tres tipos diferentes de lanzaderas sobre la distancia que alcanza un determinado misil. Los siguientes datos representan el número de millas alcanzadas por los misiles muestreados.

	Combustible 1	**Combustible 2**	**Combustible 3**
Lanzadera 1	70,4	71,7	78,5
Lanzadera 2	80,2	82,8	76,4
Lanzadera 3	90,4	85,7	84,8

Averigüe si estos datos implican que, al nivel de significación del 5%, existen diferencias en la media de millas alcanzadas cuando se utilizan:

(a) Diferentes lanzaderas

(b) Diferentes combustibles

2. Una consideración importante para decidir qué sistema de gestión de bases de datos se debe emplear es el tiempo medio requerido para aprender a utilizar el sistema. Se diseñó un contraste que afectaba a tres sistemas y a cuatro usuarios. Cada usuario necesitó los siguientes tiempos (en horas) de aprendizaje con cada sistema:

	Usario			
	1	**2**	**3**	**4**
Sistema 1	20	23	18	17
Sistema 2	20	21	17	16
Sistema 3	28	26	23	22

(a) Al nivel de significación del 5%, contraste la hipótesis de que el tiempo medio de aprendizaje es el mismo para todos los sistemas.

(b) Al nivel de significación del 5%, contraste la hipótesis de que el tiempo medio de aprendizaje es el mismo para todos los usuarios.

3. Se han plantado cinco variedades distintas de avena en cuatro superficies separadas. Las producciones resultantes fueron las siguientes.

	Superficies			
Variedad de avena	**1**	**2**	**3**	**4**
1	296	357	340	348
2	402	390	420	335
3	345	342	358	308
4	360	322	336	270
5	324	339	357	308

Averigüe si estos datos confirman la hipótesis de que la producción media no depende de:

(a) La superficie

(b) La variedad de avena

Use un nivel de significación del 5%.

4. En el ejemplo 11.3, contraste la hipótesis de que la puntuación media de un estudiante no depende del test al que se somete.

5. En el problema 1 de la sección 11.3, contraste la hipótesis de que el nivel medio de polución del aire:

(a) No cambia con el tiempo

(b) No depende del lugar

Utilice un nivel de significación del 5%.

6. En el problema 3 del la sección 11.3, contraste la hipótesis de que el número medio de cajas empaquetadas no depende de:

(a) El trabajador que hace el empaquetado

(b) El turno

Use un nivel de significación del 5%.

7. Los siguientes datos reflejan los porcentajes de fumadores en muestras aleatorias de ciudadanos del Reino Unido extraídas en distintos años.

	Edad (en años)					
Año	**16–19**	**20–24**	**25–34**	**35–49**	**50–59**	**60+**
1978	34	44	45	45	45	30
1988	28	37	36	36	33	23
1998	31	40	35	30	27	16
2000	29	35	35	29	27	16
2002	25	38	34	28	26	15

(a) Contraste la hipótesis de que los porcentajes de fumadores no dependen del año considerado.

(b) Contraste la hipótesis de que no existe efecto debido al grupo de edad.

8. En el problema 5 de la sección 11.3, contraste la hipótesis de que:

(a) Las tasas medias de natalidad no dependen del país particular considerado.

(b) Las tasas medias de natalidad no dependen del año particular considerado.

9. En el problema 7 de la sección 11.3, contraste la hipótesis de que:

(a) Las tasas medias de desempleo no dependen de la industria particular considerada.

(b) Las tasas medias de desempleo no dependen del año particular considerado.

11.5 Comentarios finales

En este capítulo se ha presentado una breve introducción de una potente técnica estadística conocida como *análisis de la varianza* (ANOVA). Esta técnica permite que los estadísticos infieran sobre las medias poblacionales cuando éstas se ven afectadas por varios factores diferentes. Pese a que solamente se han considerado problemas ANOVA con un único factor o a con dos factores, los datos de interés pueden estar afectados por un número cualquiera de factores. Adicionalmente, puede que existan interacciones entre algunos de estos factores. Por ejemplo, en el ANOVA bifactorial, podría ocurrir que la combinación de una determinada fila y una determinada columna afecte en gran medida a un valor medio. Pensemos, como caso particular, que, aunque uno solo de dos posibles genes cancerígenos pudiera ser relativamente perjudicial, la conjunción de ambos podría ser devastadora. La teoría general ANOVA muestra cómo se pueden tratar ésta y una gran variedad de situaciones.

ANOVA fue inicialmente desarrollada por R. A. Fisher, quien la aplicó a un gran número de problemas agrícolas durante su permanencia como científico jefe en los laboratorios experimentales Rothamstead. Desde entonces ANOVA se ha aplicado en múltiples áreas. Por ejemplo, en entornos de educación se podría querer analizar si el aprendizaje de álgebra por un estudiante se ve afectado por factores tales como el profesor, el programa del curso de álgebra, la duración de cada clase, el número de clases, el número de estudiantes por aula, y el texto empleado. ANOVA también se ha aplicado ampliamente en estudios de Psicología, Ciencias Sociales, Industria, Biología, y en otras muchas áreas. En verdad, ANOVA es probablemente la técnica estadística utilizada mas ampliamente.

Términos clave

Análisis de la varianza unifactorial: Modelo relativo a un conjunto de variables aleatorias. Se supone que las varianzas de estas variables son iguales y que sus valores

medios dependen de un solo factor, a saber, la muestra a la que pertenece la variable aleatoria.

Estadístico de la *F*: Un estadístico de contraste, que coincide con la razón de dos estimadores de la varianza común, cuando la hipótesis nula es cierta.

Análisis de la varianza bifactorial: Modelo en el que un conjunto de variables aleatorias normales con varianzas iguales se coloca en forma de tabla de doble entrada (con filas y columnas). La media de cualquiera de ellas depende de dos factores, a saber, la fila y la columna en la que se encuentra la variable.

Resumen

Análisis de la varianza unifactorial Consideremos m muestras independientes, cada una de tamaño n. Sean $\mu_1, \mu_2, \ldots, \mu_m$ sus respectivas medias muestrales, y si consideramos el contraste de

$$H_0\text{: todas las medias son iguales}$$

frente a

$$H_1\text{: no todas las medias son iguales}$$

Denotemos por $\overline{X}_i$ y S_i^2, respectivamente, la media muestral y la varianza muestral correspondientes a la muestra i, $i = 1, \ldots, m$. Adicionalmente, denotemos por $\overline{S}^2$ la varianza muestral del conjunto de datos $\overline{X}_1, \ldots, \overline{X}_m$.

Para contrastar H_0 *frente a* H_1, utilice el estadístico del contraste

$$\text{TS} = \frac{n\overline{S}^2}{\sum_{i=1}^{m} S_i^2/m}$$

El contraste a nivel de significación α consiste en

Rechazar H_0	si $\text{TS} \geq F_{m-1,\, m(n-1),\, \alpha}$
No rechazar H_0	en otro caso

Si el valor de TS es ν, se tiene que

$$p \text{ valor} = P\{F_{m-1,\, m(n-1)} \geq \nu\}$$

Se puede usar el programa 11-1 tanto para computar el valor de TS como para obtener el p valor resultante.

Observación: La variable $F_{r,s}$ representa una variable aleatoria F con r grados de libertad en el numerador y s en el denominador. Adicionalmente, $F_{r,s,\alpha}$ está definido de forma que

$$P\{F_{r,s} \geq F_{r,s,\alpha}\} = \alpha$$

Análisis de la varianza bifactorial

El modelo. Supongamos que cada dato viene afectado por dos factores, y que existen m posibles valores, o niveles, del primer factor y n del segundo factor. Denotemos por X_{ij} el dato obtenido cuando el primer factor está en el nivel i y el segundo factor está en el nivel j. El conjunto de datos se puede colocar en la siguiente tabla de doble entrada, por filas y columnas.

$$\begin{array}{ccccccc}
X_{11} & X_{12} & \cdots & X_{1j} & \cdots & & X_{1n} \\
X_{21} & X_{22} & \cdots & X_{2j} & \cdots & & X_{2n} \\
 & \cdots & \cdots & \cdots & \cdots & \cdots & \\
X_{i1} & X_{i2} & \cdots & X_{ij} & \cdots & & X_{in} \\
 & \cdots & \cdots & \cdots & \cdots & \cdots & \\
X_{m1} & X_{m2} & \cdots & X_{mj} & \cdots & & X_{mn}
\end{array}$$

El modelo ANOVA bifactorial supone que las X_{ij} son variables aleatorias normales con medias dadas por

$$E[X_{ij}] = \mu + \alpha_i + \beta_j$$

y varianzas comunes

$$\text{Var}\,(X_{ij}) = \sigma^2$$

Los anteriores parámetros cumplen

$$\sum_{i=1}^{m} \alpha_i = \sum_{j=1}^{n} \beta_j = 0$$

Estimación de los parámetros. Sean

$$X_{i\cdot} = \frac{\sum_{j=1}^{n} X_{ij}}{n}$$

$$X_{\cdot j} = \frac{\sum_{i=1}^{m} X_{ij}}{m}$$

$$X_{\cdot\cdot} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}}{nm}$$

Los estimadores de los parámetros son los siguientes:

$$\hat{\mu} = X_{\cdot\cdot}$$

$$\hat{\alpha}_i = X_{i\cdot} - X_{\cdot\cdot}$$

$$\hat{\beta}_j = X_{\cdot j} - X_{\cdot\cdot}$$

Contrastes de hipótesis. Sean

$$\mathrm{SS}_e = \sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} - X_{i\cdot} - X_{\cdot j} + X_{\cdot\cdot})^2$$

$$\mathrm{SS}_r = n\sum_{i=1}^{m}(X_{i\cdot} - X_{\cdot\cdot})^2$$

$$\mathrm{SS}_c = m\sum_{j=1}^{n}(X_{\cdot j} - X_{\cdot\cdot})^2$$

donde SS_e, SS_r y SS_c se denominan, respectivamente, suma de cuadrados de los errores, suma de cuadrados de las filas y suma de cuadrados de las columnas. Sea, además, $N = (n - 1)(m - 1)$.

Para contrastar H_0: *todos los* $\alpha_i = 0$ *frente a* H_1: *no todos los* $\alpha_i = 0$, utilice como estadístico del contraste

$$\mathrm{TS} = \frac{\mathrm{SS}_r/(m - 1)}{\mathrm{SS}_e/N}$$

El contraste a nivel de significación α actúa como sigue:

Rechazar H_0	si $\mathrm{TS} \geq F_{m-1,\,N,\,\alpha}$
No rechazar H_0	en otro caso

Si $\mathrm{TS} = \nu$, el p valor viene dado por

$$p \text{ valor} = P\{F_{m-1,\,N} \geq \nu\}$$

Para contrastar H_0: todos los $\beta_j = 0$ frente a H_1: no todos los $\beta_j = 0$, use el estadístico del contraste

$$\mathrm{TS} = \frac{\mathrm{SS}_c/(n - 1)}{\mathrm{SS}_e/N}$$

El contraste a nivel de significación α actúa como se indica a cotinuación:

Rechazar H_0	si $\mathrm{TS} \geq F_{n-1,\,N,\,\alpha}$
No rechazar H_0	en otro caso

Si $\mathrm{TS} = \nu$, el p valor coincide con

$$p \text{ valor} = P\{F_{n-1,\,N} \geq \nu\}$$

Puede usar el Programa 11-2 para contrastar las anteriores hipótesis. Computará los valores de los dos estadísticos del contraste y obtendrá los p valores resultantes.

Problemas de repaso

1. Una compañía dispone de tres plantas de producción aparentemente iguales. Para ver si estas plantas son igualmente efectivas, la dirección eligió aleatoriamente 30 días. Durante 10 de esos días se determinó la producción diaria en la planta 1. Durante los siguientes 10 días, se determinó la producción diaria en la planta 2, y en los 10 días finales se hizo lo mismo con la planta 3. Los siguientes datos sumariales muestran las medias muestrales y las varianzas muestrales de las unidades producidas diariamente en las tres plantas.

Planta *i*	$\overline{X}_i$	S_i^2
$i = 1$	325	450
$i = 2$	413	520
$i = 3$	366	444

Contraste la hipótesis de que el número de unidades producidas diariamente es el mismo en las tres plantas. Use un nivel de significación del 5%.

2. Se dividió aleatoriamente a 60 estudiantes que no sabían leer en cuatro grupos de 15 cada uno. A cada grupo se le dio un tipo de curso de lectura diferente. Posteriormente, los estudiantes fueron sometidos a una prueba de lectura con los resultados siguientes.

Grupo	$\overline{X}_i$	S_i^2
1	65	224
2	62	241
3	68	233
4	61	245

Contraste la hipótesis de que los cursos de lectura son igualmente efectivos. Use un nivel de significación del 5%.

3. Estudios preliminares apuntan la existencia de una posible conexión entre el color natural del pelo y el umbral de dolor. Para comprobarlo, se extrajo una muestra de 12 mujeres clasificadas por el color del pelo: claro, medio y oscuro. Cada una de ellas fue sometida a una prueba sensitiva de dolor, con los siguientes valores resultantes:

Claro	**Medio**	**Oscuro**
63	60	45
72	48	33
52	44	57
60	53	40

¿Son los datos indicados suficientes para establecer que el color del pelo afecta a los resultados de la prueba sensorial de dolor? Use un nivel de significación del 5%.

4. Se han utilizado tres lavadoras distintas para probar cuatro detergentes diferentes. Los siguientes datos muestran las puntuaciones codificadas de efectividad de cada lavadora.

	Lavadora		
Detergente	**1**	**2**	**3**
1	53	50	59
2	54	54	60
3	56	58	62
4	50	45	57

   (a) Estime la mejora en media del detergente 1 comparado (i) con el detergente 2, (ii) con el 3, y (iii) con el 4.

   (b) Estime la mejora en media de la lavadora 3 comparada (i) con la lavadora 1, y (ii) con la 2.

   (c) Contraste la hipótesis de que el detergente usado no afecta a la puntuación.

   (d) Contraste la hipótesis de que la lavadora utilizada no afecta a la puntuación.

   Tanto en (c) como en (d), utilice un nivel de significación del 5 por ciento.

5. Suponga en el problema 4 que las 12 aplicaciones de detergentes fueron todas ellas sobre lavadoras diferentes elegidas aleatoriamente. Contraste la hipótesis, al nivel de significación del 5%, de que todos los detergentes son igualmente efectivos.

6. En el ejemplo 11.3 contraste la hipótesis de que las puntuaciones medias del test dependen sólo del test y no del estudiante que se somete a él.

7. Un productor de artículos de belleza femeninos está considerando cuatro nuevas variantes de un tinte del pelo. Un elemento importante en los tintes de pelo es su poder de duración, definido como el número de días hasta que el pelo teñido no se distingue del no teñido. Para analizar el poder de duración de las nuevas variantes, la compañía contrató a tres mujeres de pelo largo. El cabello de cada mujer fue dividido en cuatro secciones, y cada sección fue tratada con un tinte distinto. Se obtuvieron los siguientes datos relativos al poder de duración de los tintes.

	Tinte			
Mujer	**1**	**2**	**3**	**4**
1	15	20	27	21
2	30	33	25	27
3	37	44	41	46

   (a) Contraste la hipótesis, al nivel de significación del 5%, de que las cuatro variantes de tinte tienen el mismo poder de duración.

   (b) Estime el poder de duración medio cuando la mujer 2 utiliza el tinte 2.

(c) Al nivel de significación del 5%, contraste la hipótesis de que el poder de duración medio no depende de la mujer que está siendo tratada.

8. Utilice los siguientes datos para contrastar la hipótesis de que (a) no existe efecto fila, y (b) no existe efecto columna.

17	23	35	39	5
42	28	19	40	14
36	23	31	44	13
27	40	25	50	17

9. El problema 9 de la sección 11.2 asume implícitamente que el número de muertes no se ve afectado por el año que se considere. Sin embargo, considere un modelo ANOVA bifactorial para este problema.

   (a) Contraste la hipótesis de que no existe efecto debido al año.

   (b) Contraste la hipótesis de que no existe efecto debido a la estación del año.

10. Los siguientes datos se refieren a las edades de mortandad de ciertas especies de ratas que fueron alimentadas con tres tipos de dietas. Las ratas elegidas eran de un tipo de corta esperanza de vida, y se dividieron aleatoriamente en tres grupos. Los datos reflejan las medias muestrales y las varianzas muestrales de las edades de mortandad (medidas en meses) de los tres grupos. Cada grupo es de tamaño 8.

	Muy baja en calorías	**Moderada en calorías**	**Alta en calorías**
Media muestral	22,4	16,8	13,7
Varianza muestral	24,0	23,2	17,1

Al nivel de significación del 5%, contraste la hipótesis de que la la dieta de las ratas no afecta a su vida media. ¿Qué ocurre si se utiliza un nivel de significación del 1%?

12

Regresión lineal

Se sabe algo cuando se entiende.

George Berkeley (filósofo británico
que dio nombre a la ciudad californiana)

Se pueden descubrir muchas cosas simplemente observando.

Yogi Berra

Se estudiará el modelo simple de regresión lineal, en el que se asume que, a menos de un error aleatorio, existe una relación lineal entre una variable de respuesta y una variable de entrada. Se utilizará el método de mínimos cuadrados para estimar los parámetros del modelo. Si se asume que el error aleatorio es normal con media 0 y varianza σ^2, se mostrará cómo contrastar hipótesis sobre los parámetros del modelo. Se introducirá el concepto de regresión a la media; se explicará cuándo surge y cómo se ha de ser cuidadoso para evitar la falacia de la regresión cuando se presente. Se explicará el coeficiente de determinación. Finalmente, se introducirá el modelo de regresión lineal múltiple, que relaciona una variable de respuesta con un conjunto de variables de entrada.

Un día de la primavera de 1888, mientras Francis Galton paseaba por el campo, meditaba sobre una cuestión que le inquietaba desde hacía ya tiempo. ¿Cuál era la relación entre las características mentales y físicas de un hijo y las de sus padres? Por ejemplo, simplificando en cierta medida sus ideas, Galton creía que la altura del hijo cuando llegará a ser adulto debería tener un valor esperado igual a la altura de su progenitor del mismo sexo. Ahora bien, si esto fuera así, se podría deducir que la mitad de la descendencia de la gente muy alta (baja) sería incluso más alta (o más baja) que sus padres. Así pues, cada nueva generación generaría individuos más altos (y también más bajos) que los de la generación anterior. Sin embargo, los datos mostraban, por el contrario, una cierta estabilidad en las alturas de la población de generación en generación. ¿Podría explicarse esta aparente contradicción?

De repente Galton vio la luz. En sus propias palabras: "Una repentina lluvia me obligó a buscar refugio en el saliente de una roca al lado del camino. De pronto, allí me surgió la idea, y el gran placer que sentí me hizo olvidar todo lo demás."

La idea de Galton fue que el valor medio de una característica del hijo (tal como la altura) no era igual al de su progenitor, sino que por el contrario estaba comprendido entre este valor y el promedio de la población completa. Como consecuencia, por ejemplo, las alturas de los descendientes de personas muy altas (llamadas por Galton personas "más altas que la mediocridad") tenderían a ser más bajas que sus padres. De igual forma, los descendientes de aquellos más bajos que la mediocridad tenderían a ser más altos que sus progenitores. Galton denominó esta idea "regresión a la mediocridad"; nosotros la conocemos como *regresión a la media*.

12.1 Introducción

A menudo, se está interesado en intentar determinar la relación que existe entre un par de variables. Por ejemplo, ¿cómo se relaciona la cantidad de dinero que se ha invertido para anunciar un nuevo producto con las cifras de ventas de dicho producto durante el primer mes? O ¿cómo se relaciona la cantidad de un catalizador empleado en un experimento científico con el resultado del experimento? O ¿está relacionada la altura de un padre con la de su hijo?

En muchas situaciones, los valores de las variables no se determinan simultáneamente en el tiempo; más bien, se ajusta una de las variables a un determinado valor, y éste, por su parte, afecta al valor de la segunda variable. Por ejemplo, el presupuesto dedicado a anuncios se suele decidir antes de que estén determinadas las cifras de ventas, y la cantidad de catalizador empleado en un experimento se suele establecer antes de que se pueda determinar el resultado del mismo. La variable cuyo valor se determina con anterioridad recibe el nombre de variable *de entrada* o variable *independiente*, mientras que a la otra se la conoce como variable *de respuesta* o *dependiente.*

Supongamos que el valor de la variable independiente se fija igual a x. Denotemos por Y el valor resultante de la variable dependiente. El tipo de relación más sencillo entre este par de variables es la relación que se establece mediante una línea recta, o relación *lineal*, en la forma

$$Y = \alpha + \beta x \tag{12.1}$$

Sin embargo, este modelo supone que (una vez que los parámetros α y β estén determinados) es posible predecir exactamente la respuesta a cualquier valor de la variable de entrada. En la práctica, tal precisión casi nunca es alcanzable, de modo que lo máximo que se puede esperar es que la anterior ecuación sea válida *sujeta a un error aleatorio.*

En la sección 12.2 se explicará con precisión el significado del modelo de *regresión lineal*, que asume que la validez de la ecuación (12.1) está sujeta a un error aleatorio. En la sección 12.3 se mostrará cómo se pueden utilizar los datos para estimar los parámetros de regresión α y β. Los estimadores que se presentan se basan en la técnica de mínimos cuadrados, que trata de encontrar la mejor recta que se ajusta a un conjunto de pares de datos. La sección 12.4 se centra sobre la variable aleatoria de *error*, la cual se supondrá que es una variable aleatoria normal con media 0 y varianza σ^2. También se considerará el problema de cómo estimarσ^2.

En la sección 12.5 se consideraran los contrastes estadísticos de la hipótesis de que no existe una relación lineal entre la variable de respuesta Y y la variable de entrada x. La sección 12.6 está dedicada a presentar el concepto de *regresión a la media*. Allí se muestra que este fenómeno aparece cuando el valor del parámetro de regresión β está comprendido entre 0 y 1. Se explicará por qué este fenómeno ocurre a menudo en situaciones de contrastación-recontrastación, y cómo un análisis poco cuidadoso de los datos nos puede llevar a la *falacia de la regresión.* Adicionalmente, se indicará en esta sección cómo se puede utilizar la regresión a la media, en conjunción con el teorema central del límite y con el paso de muchas generaciones, para explicar por qué los conjuntos de datos biológicos suelen muy habitualmente estar distribuidos normalmente.

En la sección 12.7 se tratará de encontrar un intervalo que contenga, con una determinada confianza, la respuesta futura correspondiente a cierta entrada. Estos intervalos, que utilizan los resultados obtenidos previamente, se conocen como *intervalos de predicción*. Las secciones 12.8 y 12.9 presentan, respectivamente, el coeficiente de determinación y el coeficiente de correlación. Ambos se pueden utilizar para indicar el grado de ajuste del modelo de regresión lineal a los datos. En la sección 12.10 se mostrará un método para evaluar la validez del modelo de regresión lineal, mediante el análisis de los residuos.

Finalmente, en la sección 12.11, se considerará el modelo de regresión lineal múltiple, mediante el que se intenta predecir la respuesta sin basarnos en el valor de una sola variable de entrada, sino basándonos en los valores de dos o más variables de este tipo.

12.2 Modelo de regresión lineal simple

Consideremos un par de variables, una de las cuales será denominada *variable de entrada,* y la otra, *variable de respuesta*. Supongamos que para un valor dado, x, de la variable de entrada, la variable de respuesta, Y, se puede expresar en la forma

$$Y = \alpha + \beta x + e$$

Los elementos α y β son parámetros. Se asume que la variable e, denominada *error aleatorio*, es una variable aleatoria con media 0.

Definición

La relación entre la variable de respuesta, Y, y la variable de entrada, x, especificadas ambas en la anterior ecuación, se denomina *regresión lineal simple*.

La relación de regresión lineal simple se puede expresar diciendo que para cualquier valor x de la variable de entrada, la variable de respuesta, Y, es una variable aleatoria cuya media viene dada por

$$E[Y] = \alpha + \beta x$$

Por consiguiente, un modelo de regresión lineal simple asume una relación lineal entre el valor medio de la respuesta y el valor de la variable de entrada. Los parámetros α y β serán, por lo general, desconocidos y se deberán estimar a partir de los datos.

Para ver si la regresión lineal simple se puede considerar como un modelo razonable de la relación existente entre un par de variables, se deberían recoger y representar gráficamente los datos relativos a los valores apareados de las variables. Por ejemplo, supongamos que se dispone de un conjunto de pares de datos (x_i, y_i), $i = 1, \ldots, n$, donde cada par significa que cuando la variable de entrada se ha fijado igual a x_i, el valor observado de la variable de respuesta ha sido y_i. Se deben graficar estos puntos para ver si, sujetos al error aleatorio, resulta razonable la hipótesis de que existe una relación lineal entre x e y. El gráfico citado se denomina *diagrama de dispersión*.

Ejemplo 12.1 Se ha introducido un nuevo tipo de lavadora en 11 grandes almacenes. Aproximadamente, todos los centros son de igual tamaño y están situados en comunidades similares. El productor ha variado el precio en cada gran almacén, y los datos siguientes muestran el número de unidades que se han vendido, en un mes, con los distintos precios.

Precio (en $)	Unidades vendidas
280	44
290	41
300	34
310	38
320	33
330	30
340	32
350	26
360	28
370	23
380	20

Un gráfico del número de unidades vendidas, y, frente a los precios, x, de estos 11 pares de datos viene dado en la figura 12.1. El diagrama de dispersión resultante indica que,

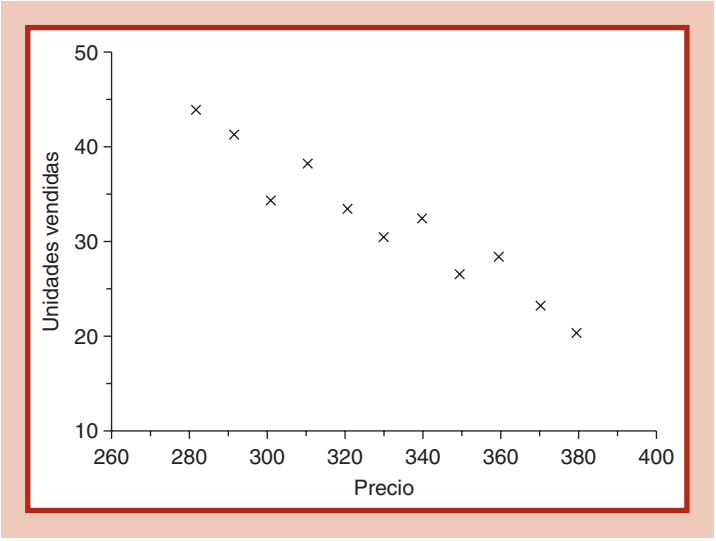


Figura 12.1 Diagrama de dispersión para los datos del ejemplo 12.1.

sujeta a errores aleatorios, la hipótesis de que existe una relación lineal entre el número de unidades vendidas y los precios resulta razonable. Esto es, el modelo de regresión lineal simple parece ser apropiado. ■

Observación: Por lo general, no se considera que la variable de entrada x en el modelo de regresión simple sea una variable aleatoria. Por el contrario, se considera que es una constante que se puede fijar a distintos valores. Por su parte, se asume que la respuesta Y es una variable aleatoria cuyo valor medio depende linealmente de la entrada x. Por esta razón se ha utilizado la letra, en mayúscula, Y para representar la respuesta. Se utiliza y, en minúscula, para denotar un valor observado de Y; así pues, y representará el valor observado de la respuesta ante un valor de entrada x.

Problemas

1. Los siguientes 12 pares de datos relacionan y, el porcentaje de producto resultante en un experimento de laboratorio, con x, la temperatura a la cual se llevó a cabo el experimento.

i	x_i	y_i	i	x_i	y_i
1	100	45	7	150	69
2	110	51	8	160	74
3	120	54	9	170	78
4	125	53	10	180	86
5	130	59	11	190	89
6	140	63	12	200	94

(a) Represente estos datos en un diagrama de dispersión.

(b) ¿Cree que el modelo de regresión simple es apropiado para describir la relación existente entre el porcentaje del producto y la temperatura?

2. Una directora de área de unos grandes almacenes desea analizar la relación entre el número de trabajadores en servicio y las pérdidas por mercancía robada. Para ello, asignó un número diferente de empleados en 10 semanas distintas. Los resultados obtenidos fueron los siguientes:

Semana	Número de trabajadores	Pérdida
1	9	420
2	11	350
3	12	360
4	13	300
5	15	225
6	18	200
7	16	230
8	14	280
9	12	315
10	10	410

(a) ¿Cuáles deberían ser las variables de entrada y de respuesta?

(b) Represente gráficamente los datos en un diagrama de dispersión.

(c) ¿Parece ser razonable el modelo de regresión lineal simple?

3. Los siguientes datos relacionan la densidad de tráfico, descrita en términos de número de automóviles por milla, con la velocidad media de tráfico en una ciudad de tamaño moderado. Los datos se recogieron en un mismo lugar y en 10 instantes distintos dentro de un periodo de 3 meses.

Densidad	Velocidad
69	25,4
56	32,5
62	28,6
119	11,3
84	21,3
74	22,1
73	22,3
90	18,5
38	37,2
22	44,6

(a) ¿Cuáles son las variables de entrada y de respuesta?

(b) Dibuje un diagrama de dispersión.

(c) ¿Parece ser razonable el modelo de regresión lineal simple?

4. Repita el problema 3, pero utilice la raíz cuadrada de la velocidad, en vez de la velocidad misma, como variable de respuesta.

5. Se sabe que la resistencia de un neumático depende de su presión. Se probó un nuevo tipo de neumático a distintas presiones, con los resultados siguientes:

Presión (en libras por pulgada)	**Resistencia (en miles de millas)**
30	29,4
31	32,2
32	35,9
33	38,4
34	36,6
35	34,8
36	35,0
37	32,2
38	30,5
39	28,6
40	27,4

(a) Dibuje un diagrama de dispersión.

(b) ¿El modelo de regresión lineal simple resulta apropiado para describir la relación existente entre la presión del neumático y su resistencia al uso?

12.3 Estimación de los parámetros de regresión

Supongamos se quieren utilizar las respuestas Y_i correspondientes a los valores de entrada x_i, $i = 1, \ldots, n$ para estimar los parámetros α y β del modelo de regresión lineal simple

$$Y = \alpha + \beta x + e$$

Para determinar los estimadores de α y β, se puede razonar como sigue: si A y B fueran los estimadores respectivos de α y β, el estimador de la respuesta correspondiente a la entrada

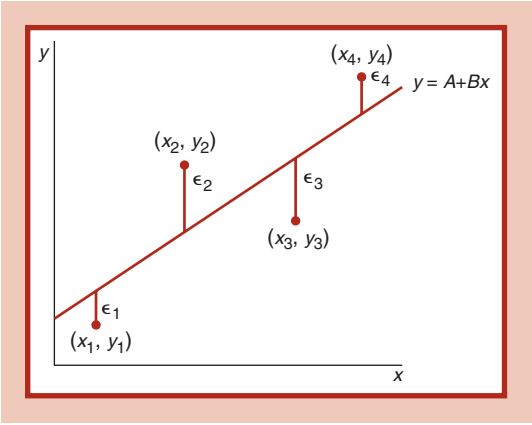


Figura 12.2 Los errores.

x_i sería $A + Bx_i$. Puesto que la respuesta observada fue Y_i, la diferencia entre la respuesta observada y su valor estimado viene dada por:

$$\epsilon_i \equiv Y_i - (A + Bx_i)$$

Es decir, ϵ_i representa el error que se deriva de usar los estimadores A y B para predecir la respuesta al valor de entrada x_i (figura 12.2).

Ahora bien, es razonable elegir como estimadores de α y β a aquellos valores A y B que hagan que estos errores sean pequeños. Para conseguir esto, se elegirán los valores de A y B que minimicen el valor de $\sum_{i=1}^{n} \epsilon_i^2$, la suma de los cuadrados de los errores. Los estimadores de α y β resultantes de este procedimiento reciben el nombre de *estimadores de mínimos cuadrados*.

Definición

Para los pares de datos dados (x_i, Y_i), $i = 1, \ldots, n$, los estimadores de (por) mínimos cuadrados son los valores de A y B que hacen

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - A - Bx_i)^2$$

lo más pequeño posible.

***Observación*:** La razón por la que se prefiere minimizar $\sum_{i=1}^{n} \epsilon_i^2$, en lugar de $\sum_{i=1}^{n} \epsilon_i$, es que la suma de los errores puede ser pequeña incluso aunque los términos de error individuales sean grandes (ya que los errores grandes positivos se podrían compensar con los errores grandes negativos). Sin embargo, esto no ocurre con la suma de los *cuadrados* de los errores, puesto que ninguno de los sumandos puede ser negativo.

Se puede demostrar que los estimadores de mínimos cuadrados de α y β, que se denotarán por $\hat{\alpha}$ y $\hat{\beta}$, vienen dados por

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}$$

donde

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{y} \quad \bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$$

La recta

$$y = \hat{\alpha} + \hat{\beta}x$$

se denomina *recta de regresión estimada*: $\hat{\beta}$ es la pendiente y $\hat{\alpha}$ es la constante (o término independiente) de la recta.

Notación: Si hacemos

$$S_{xY} = \sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})$$

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$S_{YY} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

los estimadores de mínimos cuadrados se pueden expresar mediante

$$\hat{\beta} = \frac{S_{xY}}{S_{xx}}$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}$$

Los valores de $\hat{\alpha}$ y $\hat{\beta}$ se pueden calcular a mano, o bien utilizando un ordenador. El Programa 12-1 permite computar los estimadores de mínimos cuadrados y la recta de regresión estimada. Este programa permite que el usuario calcule otros estadísticos cuyos valores se necesitarán en los apartados siguientes.

Ejemplo 12.2 Un gran banco está sopesando la conveniencia de que su personal de secretaría utilice un nuevo procesador de textos. Para conocer el periodo de aprendizaje que se necesitaría para instalar el procesador nuevo, el banco seleccionó a ocho empleados, con aproximadamente igual habilidad, para que participaran entrenamientos de duraciones distintas, y después se les puso a trabajar en un determinado proyecto. Los siguientes datos indican los tiempos de entrenamiento y los tiempos que necesitó cada empleado para realizar el proyecto (ambos en horas):

Empleado	Tiempo de entrenamiento (= x)	Tiempo para realizar el proyecto (= Y)
1	22	18,4
2	18	19,2
3	30	14,5
4	16	19,0
5	25	16,6
6	20	17,7
7	10	24,4
8	14	21,0

(a) ¿Cuál es la recta de regresión estimada?

(b) Haga una predicción del tiempo que un empleado que haya recibido un entrenamiento de 28 horas necesitaría para realizar el proyecto.

(c) Haga una predicción del tiempo que un empleado que hubiera recibido un entrenamiento de 50 horas necesitaría para realizar el proyecto.

Solución

(a) En lugar de hacer los cálculos a mano (lo cual se pide en el problema 2), se usará el Programa12-1, que computa los estimadores de mínimos cuadrados y los estadísticos asociados al modelo de regresión lineal simple. Con éste, se actúa como sigue:

Primero, introduzca el número, n, de pares de datos, que es 8.

Después, introduzca sucesivamente los 8 pares, que son:

22, 18,4
18, 19,2
30, 14,5
16, 19
25, 16,6
20, 17,7
10, 24,4
14, 21

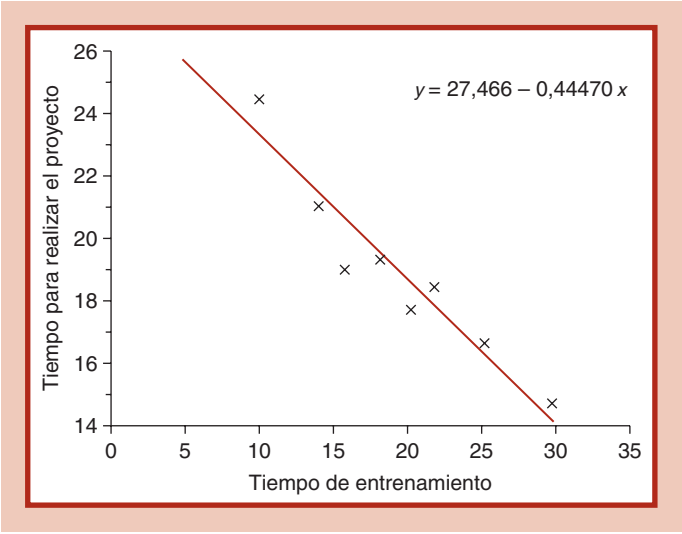


Figura 12.3 Diagrama de dispersión y línea de regresión estimada.

El programa computa los siguientes estimadores de mínimos cuadrados:

$$\begin{aligned} \mathrm{A} &= 27{,}46606 \\ \mathrm{B} &= -0{,}4447002 \end{aligned}$$

La recta de regresión estimada es la siguiente:

$$Y = 27{,}46606 - 0{,}4447002\,x$$

La figura 12.3 incluye el diagrama de dispersión y la recta de regresión resultante.

(b) La mejor predicción del tiempo de realización del proyecto correspondiente a un periodo de entrenamiento de 28 horas es su valor medio, es decir,

$$\alpha + 28\beta$$

Con los estimadores de α y β calculados previamente, la predicción del tiempo de realización será

$$27{,}466 - 28(0{,}445) = 15{,}006$$

(c) Este apartado pregunta por la predicción correspondiente al valor de entrada 50, que es mucho mayor que el resto de valores de entrada de nuestro conjunto de datos. Como resultado, aunque el diagrama de dispersión indique que el ajuste lineal proporciona una aproximación razonable dentro del rango de valores de entrada dado, uno debe ser

cauto a la hora de asumir que dicha relación continúa siendo válida para valores de entrada tan grandes como 50. Así pues, es prudente no intentar contestar al apartado (c) a partir de los datos disponibles. ■

Precaución: *No utilice la recta de regresión estimada para predecir respuestas a valores de entrada que estén muy alejados del rango de valores usado para obtener dicha recta.*

Las fórmulas siguientes pueden resultar útiles cuando los cálculos se hacen a mano.

$$S_{xY} = \sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y}$$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

Problemas

1. Calcule a mano la recta de regresión estimada con los datos siguientes:

x	*y*
1	4
2	7
3	8
5	12

   (a) Construya el diagrama de dispersión y dibuje la recta de regresión estimada.

   (b) Duplique todos los datos y repita el apartado (a).

2. Verifique los resultados del ejemplo 12.2 que se refieren a la recta de regresión estimada, haciendo los cálculos a mano o mediante una calculadora.

3. Los siguientes pares de datos representan los daños causados por incendios (en miles de dólares) en los domicilios de clase media de una gran ciudad, y las distancias (en millas) entre dichos domicilios y la estación de bomberos más cercana.

Distancia	**Daño**
5,2	36,2
4,4	28,8
3,0	22,6
1,2	8,8
7,5	41,5
9,4	25,4

(a) Dibuje un diagrama de dispersión.

(b) Intente aproximar la relación existente entre la distancia y el daño, ajustando manualmente una recta a los datos.

(c) Calcule la recta de regresión estimada y compárela con la recta dibujada en el apartado (a).

4. Considere el problema 1 de la sección 12.2.

(a) Dibuje a mano una recta que se ajuste a los datos.

(b) Determine la recta de regresión estimada y compárela con la recta dibujada en el apartado (a).

5. A continuación se muestran los consumos anuales de productos derivados del pollo (en millones de libras) en Estados Unidos, entre los años 1995 y 2002:

25,9 26,8 27,3 27,8 29,6 30,5 30,8 32,6

(a) Con el año como variable independiente y los consumos como variable dependiente, dibuje un diagrama de dispersión.

(b) Calcule la recta de regresión estimada. (Para simplificar los cálculos, se puede tomar como variable independiente el año menos 1995. Esto es, el año 1995 tomará el valor 0; 1996, el valor 1, y así sucesivamente.)

(c) Dibuje la recta de regresión estimada sobre el diagrama de dispersión.

(d) Haga una predicción del consumo correspondiente al año 1994.

(e) Haga una predicción del consumo del año 2004.

6. Los datos siguientes muestran las calificaciones medias en Matemáticas de los alumnos de ciencias del último curso de educación secundaria en una muestra de Estados, junto con el porcentaje de alumnos de eligieron la rama de ciencias.

Estado	Calificación media	Porcentaje
Arizona	525	38
California	519	54
Indiana	504	63
Missouri	583	8
Louisiana	559	8
Oregon	527	57
Virginia	510	71
Wisconsin	594	7
Texas	491	57
Vermont	512	70

Encuentre la recta de regresión estimada.

7. Los datos siguientes relacionan las producciones mundiales de pulpa de madera y de prensa escrita en siete años diferentes. Los datos provienen de la Oficina de Estadística de las Naciones Unidas, Nueva York, *Boletín Mensual de Estadística*, y aparecen en millones de toneladas métricas.

Pulpa de madera	Prensa escrita
124,4	25,4
131,3	27,8
133,1	28,3
136,6	29,3
142,0	30,6
150,1	32,3
150,3	33,1

   (a) Si la producción de pulpa de madera es la variable independiente (o de entrada), encuentre la recta de regresión estimada.

   (b) Haga una predicción de la producción de prensa escrita en un año en el que se produjeron 146,0 millones de toneladas métricas de pulpa de madera.

   (c) Si la producción de prensa escrita es la variable de entrada, encuentre la recta de regresión estimada.

   (d) Pronostique la producción mundial de pulpa de madera en un año en el que la producción de prensa escrita fue de 32,0 millones de toneladas métricas.

8. Se cree que, cuanto mayor es el contenido de alcohol en la sangre de un individuo, más lenta es su capacidad de reacción. Para comprobar esto, a 10 voluntarios se les suministraron distintas cantidades de alcohol. Tras ello, se midieron sus niveles de alcohol en la sangre, en porcentaje por unidad de peso corporal, y se midieron sus tiempos de reacción a un determinado estímulo. Los datos resultantes se muestran a continuación.

x = nivel de alcohol en la sangre (%)	y = tiempo de reacción (segundos)
0,08	0,32
0,10	0,38
0,12	0,44
0,14	0,42
0,15	0,47
0,16	0,51
0,18	0,63

(a) Dibuje un diagrama de dispersión.

(b) Aproxime la recta de regresión dibujando a mano una recta que se ajuste a los datos.

(c) ¿Cuál es la recta de regresión estimada?

(d) Compare las rectas de los apartados (b) y (c). ¿Sus pendientes son casi iguales? ¿Y sus términos independientes?

Prediga el tiempo de reacción de un individuo (no perteneciente al grupo de voluntarios) cuyo contenido de alcohol en la sangre sea:

(e) 0,15

(f) 0,17

9. En el ejemplo 12.2, supongamos que los 8 periodos de entrenamiento se fijaron por adelantado. ¿Cómo cree que se deberían haber asignado dichos periodos entre los ocho empleados?

10. En un experimento diseñado para estudiar la relación entre el número de bebidas alcohólicas consumidas y la concentración de alcohol en sangre, se asignaron aleatoriamente ciertos números de bebidas alcohólicas entre siete individuos con las mismas masas corporales. Tras una espera de 1 hora, se midieron sus niveles de alcohol en sangre. Los resultados obtenidos fueron los siguientes:

Número de bebidas	Nivel de alcohol en sangre
0,5	0,01
1	0,02
2	0,05
3	0,09
4	0,10
5	0,14
6	0,20

(a) Dibuje un diagrama de dispersión.

(b) Encuentre la recta de regresión estimada, y dibújela sobre el diagrama de dispersión.

(c) Prediga el nivel de alcohol en la sangre de una persona, con una masa corporal similar a la de los individuos del experimento, que una hora antes haya consumido 3 bebidas.

(d) ¿Qué ocurriría si la persona del apartado (c) una hora antes hubiera consumido 7 bebidas?

11. Los siguientes datos relacionan los consumos per cápita de cigarrillos en 1930 con las tasas de defunción por cáncer de pulmón en 1950, para una serie de países.

País	Consumos per cápita de cigarrillos en 1930	Defunciones por millón de habitantes en 1950
Australia	480	180
Canadá	500	150
Dinamarca	380	170
Finlandia	1100	350
Gran Bretaña	1100	460
Islandia	230	60
Holanda	490	240
Noruega	250	90
Suecia	300	110
Suiza	510	250
Estados Unidos	1300	200

(a) Determine la recta de regresión estimada.

Haga una predicción del número de defunciones por millón de habitantes debidas a cáncer de pulmón en 1950, en un país cuyo consumo de cigarrillos per cápita en 1930 fuera de:

(b) 600

(c) 850

(d) 1000

12. Los datos siguientes muestran las puntuaciones medias en Matemáticas de los estudiantes del último curso de educación secundaria en los años comprendidos entre 1980 y 1989, con la exclusión de 1983.

Año	1980	1981	1982	1984	1985	1986	1987	1988	1989
Calificación media	466	466	467	471	475	475	476	476	476

Fuente: Examen de acceso a la Universidad

(a) Haga una predicción de la calificación media de Matemáticas en 1983.

(b) Haga la misma predicción para 1993.

13. Utilice los datos del problema 3 de la sección 3.7 para predecir la puntuación en el test de inteligencia (IQ) de la hija de una mujer que ha obtenido 130 puntos IQ.

14. Utilice los datos del problema 6 de la sección 3.7 para predecir el número de adultos en libertad condicionada que hay en un Estado que tiene 14 500 adultos en prisión.

15. Los datos siguientes muestran las proporciones de mineros del carbón con síntomas de neumocomiosis y el número de años trabajados en las minas. Úselos para estimar la probabilidad de que un minero del carbón que haya trabajado 42 años padezca neumocomiosis.

Años trabajados	Proporción que padece neumocomiosis
5	0
10	0,0090
15	0,0185
20	0,0672
25	0,1542
30	0,1720
35	0,1840
40	0,2105
45	0,3570
50	0,4545

12.4 Variable aleatoria de error

Se ha definido el modelo de regresión lineal mediante la relación

$$Y = \alpha + \beta x + e$$

donde α y β son parámetros desconocidos que deben ser estimados y e es una variable aleatoria de error con media 0. Para poder hacer inferencias sobre los parámetros de regresión α y β es necesario hacer ciertas hipótesis adicionales concernientes a la variable de error e. Una hipótesis usual que se hará es que e es una variable aleatoria normal con media 0 y varianza σ^2. Así pues, se estará asumiendo que la varianza de los términos de error permanece constante con independencia de los valores de entrada x que se corresponden con los términos citados.

Dicho de otra forma, esta hipótesis equivale a asumir que, para cualquier valor de entrada x, la variable de respuesta Y es una variable aleatoria distribuida normalmente con media

$$E[Y] = \alpha + \beta x$$

y varianza

$$\text{Var}(Y) = \sigma^2$$

Otra hipótesis adicional que se hará es que todas las variables de respuesta son independientes. Esto es, se asumirá que la respuesta correspondiente, por ejemplo, a un valor de entrada x_1 es independiente de la respuesta al valor x_2.

El valor de σ^2 es desconocido y se deberá estimar a partir de los datos. Para ver cómo se consigue esto, supongamos que hemos observado las respuestas Y_i correspondientes

a los valores de entrada x_i, $i = 1, \ldots, n$. Ahora bien, para cada valor i, la variable estandarizada

$$\frac{Y_i - E[Y_i]}{\sqrt{\text{Var}(Y_i)}} = \frac{Y_i - (\alpha + \beta x_i)}{\sigma}$$

seguirá una distribución normal estándar. Así pues, dado que una variable aleatoria chi-cuadrado con n grados de libertad se define como la suma de los cuadrados de n normales estándar independientes, se sigue que

$$\frac{\sum_{i=1}^{n}(Y_i - \alpha - \beta x_i)^2}{\sigma^2}$$

es una chi-cuadrado con n grados de libertad.

Si en la expresión anterior se sustituyen α y β por sus estimadores $\hat{\alpha}$ y $\hat{\beta}$, la variable resultante continúa siendo una chi-cuadrado aunque, ahora, tiene $n - 2$ grados de libertad (puesto que se habrá perdido 1 grado de libertad por cada parámetro estimado). Esto es,

$$\frac{\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta} x_i)^2}{\sigma^2}$$

es una chi-cuadrado con $n - 2$ grados de libertad.

Las cantidades

$$Y_i - \hat{\alpha} - \hat{\beta} x_i \quad i = 1, \ldots, n$$

se denominan *residuos*. Éstos representan las diferencias entre las respuestas observadas y las predichas. Las sumas de los cuadrados de estos residuos se denotarán como SS_R. Es decir,

$$\text{SS}_R = \sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta} x_i)^2$$

Del anterior resultado se desprende

$$\frac{SS_R}{\sigma^2}$$

es una chi-cuadrado con $n - 2$ grados de libertad.

Puesto que el valor esperado de una variable aleatoria chi-cuadrado coincide con el número de sus grados de libertad, se obtiene

$$\frac{E[SS_R]}{\sigma^2} = n - 2$$

o

$$E\left[\frac{SS_R}{n - 2}\right] = \sigma^2$$

En otras palabras, se puede utilizar $SS_R/(n - 2)$ para estimar σ^2.

$$\frac{SS_R}{n - 2}$$

es el estimador de σ^2.

Se puede utilizar el Programa 12-1 para computar el valor de SS_R.

Ejemplo 12.3 Consideremos el ejemplo 12.2 y supongamos que nos interesa estimar el valor de σ^2. Para ello, se puede ejecutar el Programa 12-1, solicitándose en esta ocasión que se computen los estadísticos adicionales. Esto producirá la salida siguiente:

S(x,Y) = −125,3499
S(x,x) = 281,875
S(Y,Y) = 61,08057
SS_R = 5,337465
LA RAÍZ CUADRADA DE $(n - 2)$S(x, x)/SS_R es 17,80067

El estimador de σ^2 es 5,3375/6 = 0,8896. ■

La fórmula siguiente de SS_R resulta útil cuando se utiliza una calculadora o si se están haciendo los cálculos a mano.

La fórmula computacional para SS_R:

$$SS_R = \frac{S_{xx}S_{YY} - S_{xY}^2}{S_{xx}}$$

La forma más sencilla de calcular a mano SS_R consiste en determinar primero S_{xx}, S_{xY} y S_{YY}, para aplicar después la fórmula anterior.

Problemas

1. Estime σ^2 en el problema 1 de la sección 12.2.
2. Estime σ^2 en el problema 2 de la sección 12.2.
3. Los datos siguientes relacionan la velocidad de las pulsaciones sobre un teclado de una secretaria particular con la temperatura de su oficina. Las unidades de velocidad de pulsaciones y temperatura se dan, respectivamente, en palabras por minuto y en grados Fahrenheit.

Temperatura	Velocidad de pulsaciones
50	63
60	74
70	79

   (a) Calcule a mano el valor de SS_R.

   (b) Estime σ^2.

   (c) Si la temperatura se ha fijado en 65 grados, ¿qué velocidad de pulsaciones se puede predecir?

4. Estime σ^2 en el problema 3 de la sección 12.2.
5. Estime σ^2 en el problema 10 de la sección 12.3.
6. Los datos siguientes muestran, para determinados años comprendidos entre 1982 y 2002, los porcentajes de mujeres británicas que eran fumadoras.

Años	1982	1984	1988	1990	1994	1996	1998	2000	2002
Porcentaje	33,1	31,8	30,4	24,3	26,3	27,7	26,3	25,3	24,8

Considere que estos datos provienen de un modelo de regresión lineal, cuya entrada sea el año y la respuesta el porcentaje. Tome 1982 como año base, de modo que 1982 tenga un valor de entrada $x = 0$, 1986 tenga el valor $x = 4$, y así sucesivamente.

(a) Estime el valor de σ^2.

(b) Prediga el porcentaje de mujeres británicas que fumaban en 1997.

7. Estime σ^2 en el problema 11 de la sección 12.3.

8. Con los datos que relacionaban las edades a la que 25 padres (x) y sus respectivos hijos (Y) empezaron a afeitarse, se obtuvieron los siguientes estadísticos sumariales:

$$\bar{x} = 13{,}9 \qquad \bar{Y} = 14{,}6$$

$$S_{xx} = 46{,}8 \qquad S_{YY} = 53{,}3 \qquad S_{xY} = 12{,}2$$

(a) Determine la recta de regresión estimada.

(b) Si el padre de un muchacho se afeitó por primera vez a la edad de 15,1 años, pronostique la edad a la que comenzó a afeitarse su hijo.

(c) Estime σ^2.

12.5 Contraste de la hipótesis de que $\beta = 0$

Una hipótesis importante que se ha de considerar con respecto al modelo de regresión lineal simple

$$Y = \alpha + \beta x + e$$

es si $\beta = 0$. Su importancia estriba en el hecho de que equivale a mantener que la respuesta no depende del valor de la entrada; o, en otras palabras, no existe regresión sobre el valor de entrada.

Para contrastar

$$H_0\text{: } \beta = 0 \qquad \text{frente a} \qquad H_1\text{: } \beta \neq 0$$

es necesario primero estudiar la distribución de $\hat{\beta}$, el estimador de β. Es decir, se querrá rechazar H_0 cuando $\hat{\beta}$ se encuentre alejado de 0, y no se rechazará en otro caso. Para determinar la separación que debe existir entre $\hat{\beta}$ y 0 para que esté justificado el rechazo de la hipótesis nula, es necesario conocer su distribución.

Se puede demostrar que $\hat{\beta}$ se distribuye normalmente con media y varianza, respectivas, dadas por

$$E[\hat{\beta}] = \beta$$

y

$$\text{Var}(\hat{\beta}) = \frac{\sigma^2}{S_{xx}}$$

De donde, la variable estandarizada

$$\frac{\hat{\beta} - \beta}{\sqrt{\sigma^2/S_{xx}}} = \sqrt{\frac{S_{xx}}{\sigma^2}}(\hat{\beta} - \beta)$$

sigue una distribución normal estándar.

Sin embargo, el contraste no se puede basar directamente en el resultado anterior, ya que la variable estandarizada involucra el parámetro desconocido σ^2. Pese a ello, se puede demostrar que, si se reemplaza σ^2 por su estimador $SS_R/(n-2)$, que es una chi-cuadrado con $n-2$ grados de libertad, el resultado sigue una distribución t con $n-2$ grados de libertad. Esto es,

$$\sqrt{\frac{(n-2)S_{xx}}{SS_R}}(\hat{\beta} - \beta)$$

sigue una distribución t con $n-2$ grados de libertad.

De lo anterior se obtiene que si H_0 es cierta y, por tanto, $\beta = 0$,

$$\sqrt{\frac{(n-2)S_{xx}}{SS_R}}\hat{\beta}$$

sigue una distribución t con $n-2$ grados de libertad. Esto da pié al siguiente contraste de H_0.

El contraste, al nivel de significación γ, consiste en

Rechazar H_0	si $\quad \lvert TS\rvert \geq t_{n-2,\,\gamma/2}$
No rechazar H_0	en otro caso

donde

$$TS = \sqrt{\frac{(n-2)S_{xx}}{SS_R}}\hat{\beta}$$

Una forma equivalente de llevar a cabo el contraste consiste en computar primero el valor del estadístico del contraste, TS; asumamos que éste es ν. La hipótesis nula se debe

rechazar si el nivel de significación fijado, γ, es al menos tan grande como el p valor dado por

$$p \text{ valor } = P\{|T_{n-2}| \geq |\gamma|\}$$
$$= 2P\{T_{n-2} \geq |\gamma|\}$$

donde T_{n-2} es una variable aleatoria t con $n - 2$ grados de libertad. Se puede utilizar el Programa 8-2 para calcular esta última probabilidad.

Ejemplo 12.4 Un individuo mantiene que el consumo de gasolina de su automóvil no depende de la velocidad del vehículo. Para contrastar la veracidad de esta hipótesis, se probó el automóvil circulando a distintas velocidades entre 45 y 75 millas por hora. A continuación, se muestran las millas por galón de carburante que se consiguieron con las velocidades citadas.

Velocidad	Millas por galón
45	24,2
50	25,0
55	23,3
60	22,0
65	21,5
70	20,6
75	19,8

¿Niegan estos datos la idea de que la velocidad de circulación del vehículo no afecta al número de millas recorridas por galón de carburante?

Solución Supongamos que el modelo de regresión simple

$$Y = \alpha + \beta x + e$$

relaciona Y, las millas recorridas por galón, con x, la velocidad a la cual circula el vehículo. La idea del conductor coincide con el hecho de que el coeficiente de regresión β es igual a 0. Para ver si los datos tienen la fuerza suficiente para negar esta idea es necesario comprobar si nos conducen a rechazar la hipótesis nula en el contraste de

$$H_0: \beta = 0 \qquad \text{frente a} \qquad H_1: \beta \neq 0$$

Para calcular el valor del estadístico del contraste, primero, se computarán los valores de S_{xx}, S_{YY} y S_{xY}. Mediante cálculos manuales se obtiene que

$$S_{xx} = 700 \qquad S_{YY} = 21{,}757 \qquad S_{xY} = -119$$

Con la fórmula presentada al final de la sección 12.4 para calcular SS_R se obtiene

$$SS_R = \frac{S_{xx}S_{YY} - S_{xY}^2}{S_{xx}}$$

$$= \frac{700(21{,}757) - 119^2}{700} = 1{,}527$$

De donde

$$\hat{\beta} = \frac{S_{xY}}{S_{xx}} = \frac{-119}{700} = -0{,}17$$

El valor del estadístico del contraste es

$$TS = \sqrt{\frac{5(700)}{1{,}527}}(-0{,}17) = -8{,}139$$

Puesto que, de la tabla D.2 del Apéndice, $t_{5,\,0{,}005} = 4{,}032$, se sigue que la hipótesis de que $\beta = 0$ se debe rechazar al nivel de significación del 1%. Así pues, igualmente, se ha debe rechazar la conjetura de que las millas recorridas por galón no dependen de la velocidad de circulación del vehículo. De hecho, está claro que resulta bastante evidente que un aumento en la velocidad produce una disminución en las distancias recorridas por galón. ■

Problemas

1. Contraste la hipótesis de que $\beta = 0$ con los datos siguientes:

x	Y
3	7
8	8
10	6
13	7

Utilice un nivel de significación del 5%.

2. El conjunto de datos siguiente muestra las alturas de 12 antiguos estudiantes de la Facultad de Derecho cuyas calificaciones en la carrera fueron más o menos similares. También muestra sus salarios 5 años después de su graduación. Todos ellos ejercieron como abogados. Las alturas se han medido en pulgadas y los salarios en unidades de 1000 dólares.

Altura	Salario
64	111
65	114
66	108
67	123
69	97
70	116
72	125
72	108
74	142
74	122
75	110
76	134

(a) ¿Establecen estos datos que los salarios de los abogados están relacionados con sus alturas?

(b) ¿Cuál es la hipótesis nula en el apartado (a)?

3. La tabla siguiente relaciona el número de manchas solares que aparecieron cada año desde 1970 hasta 1980 con el número de fallecimientos por accidentes de automóvil en dichos años. Los datos sobre fallecimientos por accidentes de tráfico aparecen en unidades de 1000 fallecidos.

Año	Manchas solares	Fallecimientos por accidente de tráfico
70	165	54,6
71	89	53,3
72	55	56,3
73	34	49,6
74	9	47,1
75	30	45,9
76	59	48,5
77	83	50,1
78	109	52,4
79	127	52,5
80	153	53,2

Contraste la hipótesis de que el número de fallecimientos por accidentes de tráfico no guarda relación con el número de manchas solares. Utilice un nivel de significación del 5%.

4. Una compañía de electricidad pretende estimar la relación existente entre la temperatura diaria en verano y la cantidad de electricidad demandada por sus usuarios. Se recogieron los siguientes datos:

Temperatura (en grados Fahrenheit)	Electricidad (en millones de kilowatios)
85	22,5
90	23,7
76	20,3
91	23,4
84	24,2
94	23,5
88	22,9
85	22,4
97	26,1
86	23,1
82	22,5
78	20,9
77	21,0
83	22,6

(a) Encuentre la recta de regresión estimada.

(b) Prediga la electricidad que se consumiría si la temperatura de mañana fuera de 93 grados.

(c) Contraste la hipótesis de que la temperatura diaria no afecta al consumo de electricidad, al nivel de significación del 5%.

Los problemas del 5 al 8 se refrieren a los datos siguientes, que relacionan el consumo de cigarrillos con las tasas de mortalidad por cuatro tipos de cáncer en 14 Estados. Los datos se basan parcialmente en registros de los impuestos sobre el tabaco relativos al año 1960.

Consumo de cigarrillos y tasas de mortalidad por cáncer

		Muertes anuales por 100 000 habitantes			
Estado	Cigarrillos por persona	Cáncer de vejiga	Cáncer de pulmón	Cáncer de riñón	Leucemia
California	2860	4,46	22,07	2,66	7,06
Idaho	2010	3,08	13,58	2,46	6,62
Illinois	2791	4,75	22,80	2,95	7,27
Indiana	2618	4,09	20,30	2,81	7,00
Iowa	2212	4,23	16,59	2,90	7,69
Kansas	2184	2,91	16,84	2,88	7,42
Kentucky	2344	2,86	17,71	2,13	6,41
Massachusetts	2692	4,69	22,04	3,03	6,89
Minnesota	2206	3,72	14,20	3,54	8,28

Consumo de cigarrillos y tasas de mortalidad por cáncer (*Continuación*)

		Muertes anuales por 100 000 habitantes			
Estado	**Cigarrillos por persona**	**Cáncer de vejiga**	**Cáncer de pulmón**	**Cáncer de riñón**	**Leucemia**
New York	2914	5,30	25,02	3,10	7,23
Alaska	3034	3,46	25,88	4,32	4,90
Nevada	4240	6,54	23,03	2,85	6,67
Utah	1400	3,31	12,01	2,20	6,71
Texas	2257	3,21	20,74	2,69	7,02

5. (a) Dibuje un diagrama de dispersión de los consumos de cigarrillos frente a la tasa de mortalidad por cáncer de vejiga.

   (b) Encuentre la recta de regresión estimada.

   (c) Contraste la hipótesis, al nivel de significación del 5%, de que el consumo de cigarrillos no afecta a la tasa de mortalidad por cáncer de vejiga.

   (d) Repita el apartado (c) al nivel de significación del 1%.

6. (a) Dibuje un diagrama de dispersión del consumo de cigarrillos frente a la tasa de mortalidad por cáncer de pulmón.

   (b) Encuentre la recta de regresión estimada.

   (c) Contraste la hipótesis de que el consumo de cigarrillos no afecta a la tasa de mortalidad por cáncer de pulmón, al nivel de significación del 5%.

   (d) Repita el apartado (c) al nivel de significación del 1%.

7. (a) Dibuje un diagrama de dispersión del consumo de cigarrillos frente a la tasa de mortalidad por cáncer de riñón.

   (b) Encuentre la recta de regresión estimada.

   (c) Contraste la hipótesis de que el consumo de cigarrillos no afecta a la tasa de mortalidad por cáncer de riñón, al nivel de significación del 5%.

   (d) Repita el apartado (c) al nivel de significación del 1%.

8. (a) Dibuje un diagrama de dispersión del consumo de cigarrillos frente a la tasa de mortalidad por leucemia.

   (b) Encuentre la recta de regresión estimada.

   (c) Contraste la hipótesis de que el consumo de cigarrillos no afecta a la tasa de mortalidad por leucemia, al nivel de significación del 5%.

   (d) Repita el apartado (c) al nivel de significación del 1%.

9. En el problema 3 de la sección 12.3, contraste la hipótesis de que los daños por incendios no dependen de las distancias a la estación de bomberos más próxima. Utilice un nivel de significación del 5%.

10. La tabla siguiente muestra el porcentaje de chicos y chicas británicos de 15 años de edad que eran fumadores, en una muestra de los años comprendidos entre 1982 y 2003. Úsela para:

    (a) Contrastar la hipótesis, al nivel de significación del 5%, de que el porcentaje de chicos que fuman se mantuvo constante en el tiempo.

    (b) Contrastar la hipótesis, al nivel de significación del 5%, de que el porcentaje de chicas fumadoras se mantuvo constante en el tiempo.

    (c) Contrastar la hipótesis, al nivel de significación del 5%, de que el porcentaje de personas de 15 años que fuman se mantuvo constante en el tiempo.

Porcentaje de alumnos de 15 años que son fumadores habituales (al menos 1 cigarrillo/semana en media), Inglaterra

	1982	1984	1986	1988	1990	1992	1994	1996	1998	2000	2003
Chicos	24	28	18	17	25	21	26	28	19	21	18
Chicas	25	28	27	22	25	25	30	33	29	26	26
Total	25	28	22	20	25	23	28	30	24	23	22

11. La tabla siguiente muestra el consumo per cápita de plátanos, manzanas y naranjas (en libras) en Estados Unidos durante siete años distintos.

Año	Plátanos	Manzanas	Naranjas	Año	Plátanos	Manzanas	Naranjas
1970	17,4	16,2	15,7	1983	21,2	17,6	15,6
1975	17,6	18,2	15,4	1985	23,4	16,6	12,0
1980	20,8	18,3	15,4	1987	24,9	20,3	13,9
1982	22,5	17,1	12,3				

Fuente: Departamento de Agricultura de Estados Unidos, *Consumo de alimentos, precios y gastos.*

Contraste la hipótesis de que el consumo anual de plátanos no depende de las cantidades anuales de:

(a) Manzanas que se han consumido.

(b) Naranjas que se han consumido.

(c) Contraste la hipótesis de que los consumos anuales per cápita de naranjas no guarda relación con los consumos anuales de manzanas.

12.6 Regresión a la media

El término *regresión* fue inicialmente empleado por Francis Galton en sus estudios sobre las leyes de la herencia. Galton mantuvo que estas leyes ocasionaban que los valores extremos sobre una población *"regresaran a la media"*. Por ello, él entendía que los descendientes de individuos que tenían valores extremos con respecto a cierta característica tendían a tener menores valores extremos que sus padres.

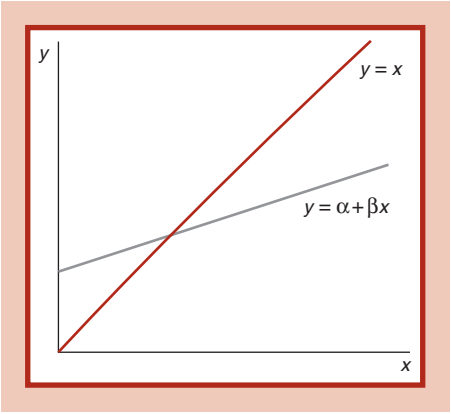


Figura 12.4 La regresión a la media ocurre cuando $0 < \beta < 1$. Para valores de x pequeños, $\alpha + \beta x > x$; para valores de x grandes, $\alpha + \beta x < x$.

Si se asume un modelo de regresión lineal entre la característica de los descendientes, Y, y la de los padres, x, la regresión a la media ocurrirá cuando el parámetro de regresión β se encuentre entre 0 y 1. Es decir, si

$$E[Y] = \alpha + \beta x$$

y $0 < \beta < 1$, $E[Y]$ será menor que x cuando x sea grande, y será menor que x cuando x sea pequeño. Que esta aserción es cierta se puede comprobar algebraicamente o bien con una representación gráfica de las dos rectas

$$y = \alpha + \beta x$$

y

$$y = x$$

El gráfico indica que, cuando $0 < \beta < 1$, la recta $y = \alpha + \beta x$ está por encima de la recta $y = x$ para valores reducidos de x, y que está por debajo para valores altos de x. La figura 12.4 representa este gráfico.

Ejemplo 12.5 Para ilustrar la tesis de Galton sobre la regresión a la media, el estadístico británico Karl Pearson hizo una gráfica de las alturas de 10 hijos elegidos aleatoriamente frente a las de sus padres. Los datos representados gráficamente (en pulgadas) fueron los siguientes.

Altura del padre	Altura del hijo	Altura del padre	Altura del hijo
60	63,6	67	67,1
62	65,2	68	67,4
64	66	70	68,3
65	65,5	72	70,1
66	66,9	74	70

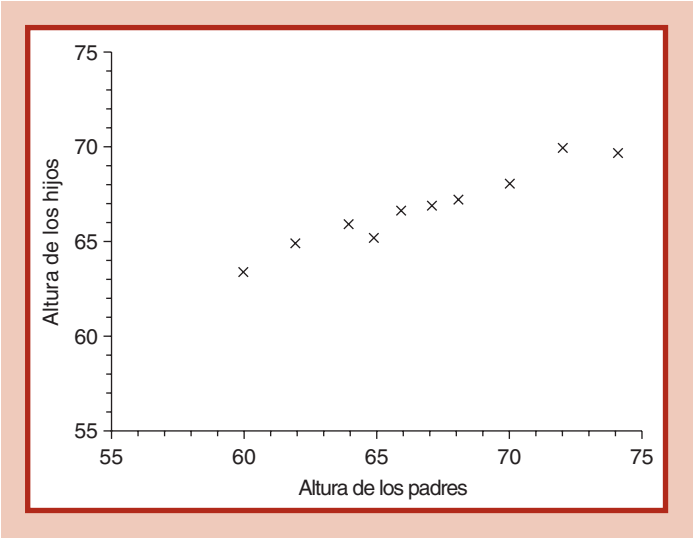


Figura 12.5 Diagrama de dispersión de las alturas de los hijos frente a las de los padres.

Un diagrama de dispersión que representa estos datos se muestra en la figura 12.5.

Observe que, aunque los datos parecen indicar que los padres más altos tienden a tener hijos más altos, también indican que los hijos de padres que son extremadamente altos o extremadamente bajos tienden a aproximarse a la "media" más que sus padres; es decir, se produce una *regresión a la media.*

A continuación se determinará si los datos precedentes tienen la fuerza suficiente para probar que existe una regresión a la media. Para ello, se tomará este último aserto como hipótesis alternativa. Esto es, utilizaremos los datos para contrastar

$$H_0: \beta \geq 1 \qquad \text{frente a} \qquad H_1: \beta < 1$$

Ahora bien, este contraste es equivalente a contrastar

$$H_0: \beta = 1 \qquad \text{frente a} \qquad H_1: \beta < 1$$

y nos basaremos en el hecho de que

$$\sqrt{\frac{(n-2)S_{xx}}{SS_R}}\,(\hat{\beta} - \beta)$$

sigue una distribución t con $n - 2$ grados de libertad.

De donde, cuando $\beta = 1$, el estadístico del contraste

$$\text{TS} = \sqrt{\frac{8S_{xx}}{\text{SS}_R}}(\hat{\beta} - 1)$$

sigue una distribución t con 8 grados de libertad. El contraste, a nivel de significación α, consistirá en rechazar H_0 cuando el valor de TS sea suficientemente pequeño (ya que esto ocurrirá cuando $\hat{\beta}$, el estimador de β, sea suficientemente menor que 1). Específicamente, el contraste consistirá en:

Rechazar H_0	si $\text{TS} \leq -t_{8,\alpha}$
No rechazar H_0	en otro caso

Para determinar el valor del estadístico del contraste, TS, se puede ejecutar el Programa 12-1, con el que se obtendrá:

Los estimadores de mínimos cuadrados son los siguientes
A = 35,97757
B = 0,4645573

La recta de regresión estimada es
Y = 35,97757 + 0,4645573x
S(x,Y) = 79,71875
S(x,x) = 171,6016
S(Y,Y) = 38,53125
SS_R = 1,497325

La raíz cuadrada de $(n - 2)\text{S(x, x)}/\text{SS}_R$ es 30,27942

De lo anterior se ve que

$$\text{TS} = 30{,}2794(0{,}4646 - 1) = -16{,}21$$

Puesto que $t_{8,\,0{,}01} = 2{,}896$, resulta que

$$\text{TS} < -t_{8,\,0{,}01}$$

y, por tanto, la hipótesis nula de que $\beta \geq 1$ se debe rechazar al nivel de significación del 1%. De hecho, el p valor es

$$p \text{ valor} = P\{T_8 \leq -16{,}213\} \approx 0$$

por lo que se rechazará la hipótesis nula de que $\beta \geq 1$ a (casi) cualquier nivel de significación. De esta forma, se ha comprobado la regresión a la media. ■

Una explicación biológica moderna del fenómeno de la regresión a la media se basa, grosso modo, en la consideración de que, puesto que un descendiente obtiene una selección aleatoria de la mitad de los genes de cada uno de sus padres, un descendiente con uno de sus progenitores muy alto tendería a tener genes menos "altos" que los de dicho progenitor.

Aunque las aplicaciones más importantes del fenómeno de la regresión a la media afectan a la relación entre las características biológicas de un descendiente con respecto a las de sus progenitores, este fenómeno también surge en situaciones en las que se dispone de dos conjuntos de datos que se refieren las mismas variables. Esto se ilustra en el ejemplo 12.6.

Ejemplo 12.6 Los datos siguientes relacionan el número de fallecimientos por accidentes de tráfico ocurridos en 12 condados del noroeste de Estados Unidos en los años 1988 y 1989.

Condado	Fallecimientos en 1988	Fallecimientos en 1989
1	121	104
2	96	91
3	85	101
4	113	110
5	102	117
6	118	108
7	90	96
8	84	102
9	107	114
10	112	96
11	95	88
12	101	106

El diagrama de dispersión de estos datos se muestra en la figura 12.6. Una ojeada a esta figura indica que en 1989 se produjo una reducción en la mayor parte de los condados que tuvieron un alto número de fallecimientos en 1988. De igual forma, aparentemente se produjo un incremento en aquellos condados con valores bajos en 1988. En consecuencia, se puede intuir que, en efecto, se ha producido una regresión a la media. Con el Programa 12-1 se puede obtener la ecuación de regresión estimada

$$y = 74{,}589 + 0{,}276x$$

la cual muestra que el valor estimado de β parece ser realmente menor que 1.

Se ha de ser cuidadoso cuando se considera la razón que sustenta el fenómeno de regresión a la media con los datos precedentes. Por ejemplo, podría ser natural suponer que en aquellos condados con un alto número de fallecimientos por accidentes de tráfico en 1988

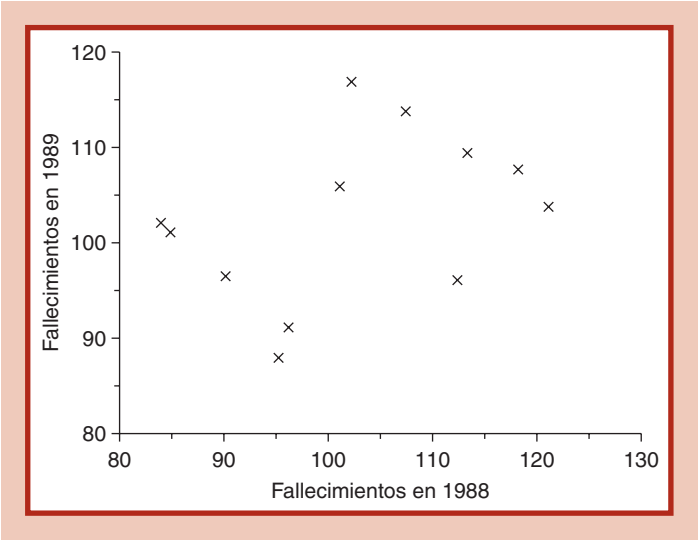


Figura 12.6 Diagrama de dispersión de los fallecimientos que hubo en 1989 frente a los de 1988.

han hecho un gran esfuerzo –quizás mejorando la seguridad de sus carreteras o haciendo que la gente sea más consciente de los potenciales peligros derivados de una conducción insegura– para reducir este número. Adicionalmente, se podría suponer que en aquellos condados con un número menor de fallecimientos en 1988 hayan relajado su atención y no se hayan esforzado en conseguir posteriores mejoras en sus cifras, y, como resultado, se habría incrementado el número de fallecimientos en el año siguiente.

Aunque las suposiciones anteriores podrían ser correctas, es importante tener en cuenta que la regresión a la media se podría haber producido sin que ninguno de los condados hubiera realizado nada extraordinario. De hecho, podría realmente ocurrir que los condados con un alto número de fallecimientos en 1988 fueran muy desafortunados dicho año, y que, por tanto, una disminución en sus cifras del año próximo reflejara simplemente que han tenido un retorno a la normalidad. (Por analogía, si resultaran 9 caras en 10 lanzamientos de una moneda no sesgada, es bastante probable que en 10 posteriores lanzamientos de la moneda resultasen menos de 9 caras.) Similarmente, aquellos países con pocos fallecimientos en 1988 podrían haber tenido suerte en dicho año, y un resultado más normal en 1989 justificaría un aumento.

La creencia equivocada de que la regresión a la media se debe a alguna influencia exterior, cuando en realidad se ha debido al azar, se denotada con frecuencia como la *falacia de la regresión.* ■

La regresión a la media juega un papel clave para explicar por qué tantos conjuntos de datos biológicos en poblaciones homogéneas tienden a seguir una distribución normal. Por ejemplo, si se representaran gráficamente las alturas de todas las chicas que asisten a un curso de secundaria, sería una buena apuesta mantener que el histograma resultante tendría una

forma muy parecida a la curva normal. Una posible explicación de este hecho combina el teorema central del límite, la regresión a la media y el paso de un gran número de generaciones. Se esbozará a continuación la explicación citada.

*12.6.1 Por qué los conjuntos de datos biológicos con frecuencia se distribuyen normalmente

Se presentará este argumento en el contexto del análisis de las alturas de las mujeres de una población. Se observará la evolución de esta población de mujeres a lo largo de un gran número de generaciones. Supongamos que inicialmente existen k mujeres, a las que denominaremos *generación inicial*, cuyas alturas son $x_1, \ldots, x_k$. Estos valores son totalmente arbitrarios. Denotemos como d a la diferencia entre el mayor y el menor de estos valores. Por ejemplo, si

$$k = 3 \qquad x_1 = 60 \qquad x_2 = 58 \qquad x_3 = 66$$

se tendrá que $d = 66 - 58 = 8$.

Si Y denota la altura de una hija de una mujer cuya altura es x, se asumirá el modelo de regresión lineal

$$Y = \alpha + \beta x + e$$

En este modelo se supondrá, como es habitual, que la variable aleatoria de error se distribuye normalmente con media 0 y varianza σ^2. Sin embargo, mientras que esta hipótesis se asume a menudo sin ninguna justificación, en nuestra aplicación resulta bastante razonable debido al teorema central del límite. Esto es, la altura de una hija de una mujer que mide x se puede concebir como la suma de un gran número de variables aleatorias aproximadamente independientes que dependen, entre otros factores, tanto del conjunto aleatorio de los genes que recibe como de los factores ambientales. De esto se desprende, por el teorema central del límite, que su altura deberá seguir aproximadamente una distribución normal. Supongamos, además, que efectivamente se produce una regresión a la media, es decir, $0 < \beta < 1$.

Se tendrá, pues, que todas las alturas de las hijas de las k mujeres de la población inicial se distribuyen normalmente. Sin embargo, es importante observar que sus alturas medias son todas distintas. Por ejemplo, la hija de una mujer de altura x_1 tendrá una altura distribuida normalmente con una media $\alpha + \beta x_1$, mientras que la hija de una mujer de altura x_2 tendrá una altura media diferente a la anterior e igual a $\alpha + \beta x_2$. En consecuencia, las alturas de todas las hijas no provienen de la *misma* distribución normal y, por esta razón, un gráfico de todas sus alturas no tiene por qué seguir una curva normal.

Sin embargo, si consideramos la diferencia entre la mayor y la menor altura media de todas las hijas del conjunto inicial de mujeres, no es difícil comprobar que

$$\text{Diferencia} \leq \beta d$$

(Si cada una de las mujeres del conjunto inicial tuviera al menos una hija, esta desigualdad se convertiría en una igualdad.) Si ahora considerásemos a las hijas de estas hijas, se puede

demostrar que sus alturas se distribuirían normalmente con medias diferentes y con varianzas iguales. Igualmente, se puede comprobar que la diferencia entre la mayor y la menor altura media de estas hijas de la segunda generación verifica que

$$\text{Diferencia} \leq \beta^2 d$$

De hecho, a medida que pasan más y más generaciones, si considerásemos a las hijas de la n-ésima generación posterior a la inicial, se podría comprobar que las alturas de las mujeres de la generación citada se distribuyen normalmente con varianzas iguales y con medias que, aunque distintas, son tales que la diferencia entre la mayor y la menor de ellas satisface

$$\text{Diferencia} \leq \beta^n d$$

Ahora bien, puesto que $0 < \beta < 1$, se tiene que a medida que n crece, $\beta^n d$ se aproxima cada vez más a 0. Así pues, transcurrido un número de generaciones suficientemente grande, todas las mujeres de la población tendrían unas alturas distribuidas normalmente con aproximadamente la misma media e igual varianza. Esto es, tras un alto número de generaciones, las alturas de las mujeres provendrán aproximadamente de la misma población normal y, por consiguiente, un gráfico de estas alturas debería reflejar la forma acampanada de la curva normal.

Problemas

1. Los datos siguientes provienen de un experimento llevado a cabo por Francis Galton. Los datos relacionan el diámetro de una semilla descendiente con el de su semilla progenitora, en un caso de autofertilización.

Diámetro de la semilla progenitora	Diámetro de la la semilla descendiente
15	15,3
16	16,4
17	15,5
18	16,2
19	16,0
20	17,4
21	17,5

(a) Estime los parámetros de regresión.

(b) ¿Aparenta existir una regresión a la media?

2. En el ejemplo 12.6 se comprobó que el valor de β es menor que 1. Con los datos de este ejemplo, contraste las hipótesis

$$H_0: \beta \geq 1 \qquad \text{frente a} \qquad H_1: \beta < 1$$

¿Se debería rechazar H_0 al nivel de significación del 5%?

3. ¿Sería sorprendente que los siguientes conjuntos de datos exhibieran una regresión a la media? ¿Se puede esperar que efectivamente muestren este fenómeno? Explique sus respuestas.

   (a) Una persona va a 10 restaurantes diferentes, totalmente desconocidos para él. Come una vez en cada uno de ellos y, según la calidad de la comida, le otorga una puntuación comprendida entre 0 y 100. Posteriormente, vuelve a esos restaurantes y vuelve a poner una calificación a la comida recibida. Los datos son las dos puntuaciones dadas a cada uno de los 10 restaurantes.

   (b) En un determinado momento, se mide el latido del corazón, en número de pulsaciones por segundo, de 12 individuos. Denotemos estos valores como x. Una hora más tarde, se repiten las mediciones para obtener los valores y.

   (c) El conjunto considera datos apareados relativos a distintos fondos de inversión. Para cada fondo, la variable x indica la clasificación del fondo en 1995, mientras que la variable y refleja la misma clasificación en 1996.

   (d) Los datos representan las calificaciones obtenidas por 20 alumnos en el primer curso de preescolar en dos pruebas de un mismo test de inteligencia (IQ). La primera se hizo la primera semana del curso y la segunda, un mes más tarde.

4. Contraste

$$H_0: \beta = 1 \qquad \text{frente a} \qquad H_1: \beta < 1$$

con el siguiente conjunto de datos. Utilice un nivel de significación del 5%.

x	y
24	27
21	24
26	20
17	22
15	21
24	20
23	17

5. Los siguientes datos muestran las calificaciones medias en Matemáticas de los estudiantes que se han presentado en un examen de acceso a la universidad en los años 2000 y 2002 para una muestra de diferentes Estados.

Estado	2000	2002
Arizona	523	523
California	518	517
Indiana	501	503
Missouri	577	580
Florida	486	473
Oregon	527	528
Virginia	500	506
Wisconsin	597	599
Texas	500	500
Vermont	508	510

   (a) Encuentre la recta de regresión estimada.

   (b) ¿Muestran los datos una regresión a la media?

6. La figura 12.7 presenta un histograma de las alturas (en pulgadas) de 8585 hombres. ¿Cuál es su bondad de ajuste a la curva normal?

7. La figura 12.8 muestra un histograma de los pesos de 7738 hombres. ¿Cuál es su bondad de ajuste a la curva normal?

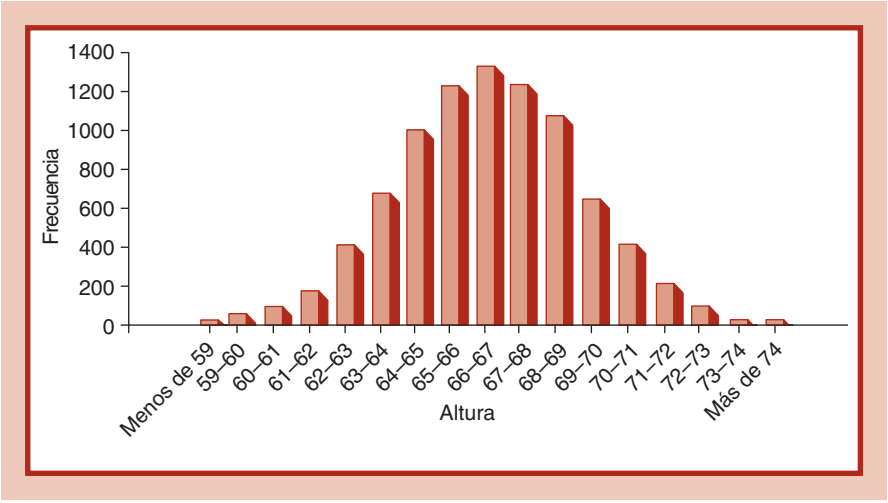


Figura 12.7 Histograma de alturas.

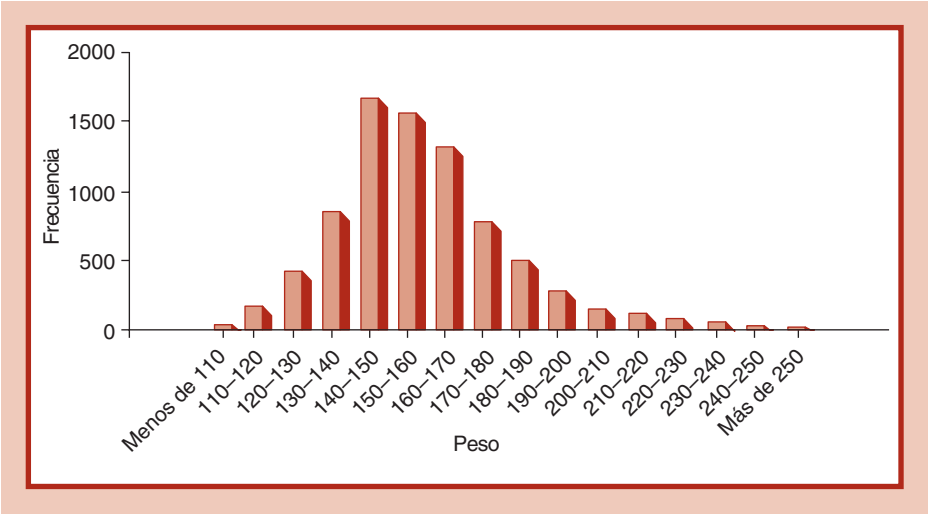


Figura 12.8 Histograma de pesos.

12.7 Intervalos de predicción para respuestas futuras

Supongamos, en un modelo de regresión lineal, que los valores de entrada x_i han producido los valores de respuesta y_i, $i = 1, \ldots, n$. La mejor predicción de la respuesta a una nueva entrada x_0 es, naturalmente, $\hat{\alpha} + \hat{\beta}x_0$. Sin embargo, en vez de predecir con un simple número, suele ser más útil obtener un intervalo de predicción que contenga el valor de la respuesta con cierto grado de confianza. El citado intervalo de predicción viene dado a continuación.

Intervalo de predicción para la respuesta a una entrada x_0, basado en los valores de respuesta y_i correspondientes a los valores de entrada x_i, $i = 1, \ldots, n$:

Con un nivel de confianza $100(1 - \gamma)$, la respuesta Y al valor de entrada x_0 cae en el intervalo

$$\hat{\alpha} + \hat{\beta}x_0 \pm t_{n-2,\,\gamma/2}W$$

donde $t_{n-2,\,\gamma/2}$ denota el percentil de orden $100(1 - \gamma/2)$ por ciento de la distribución t con $n - 2$ grados de libertad, y

$$W = \sqrt{\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]\frac{\mathrm{SS}_R}{n-2}}$$

Todos los valores $\hat{\alpha}$, $\hat{\beta}$, $\bar{x}$, S_{xx}, y SS_R se calculan a partir de los datos x_i, y_i, $i = 1, \ldots, n$.

Ejemplo 12.7 Con los datos del ejemplo 12.6, determine, con una confianza del 95%, un intervalo que contenga la altura, una vez adulto, de un hijo recién nacido cuyo padre mida 70 pulgadas.

Solución De la salida del programa 12-1 se obtiene

$$\hat{\alpha} + 70\hat{\beta} = 68{,}497$$
$$W = 0{,}4659$$

Puesto que la tabla D.2 indica que $t_{8,\,0{,}025} = 2{,}306$, se ve que el intervalo de predicción, al 95% de confianza, para la altura de un hijo cuyo padre mida 70 pulgadas es

$$68{,}497 \pm 2{,}306(0{,}4659) = 68{,}497 \pm 1{,}074$$

Esto es, con una confianza del 95%, se tendrá que la altura del hijo, una vez adulto, estará comprendida entre 67,423 y 69,571 pulgadas. ■

Ejemplo 12.8 Una compañía que regenta una concesión de hamburguesas en un estadio de fútbol universitario tiene que decidir el lunes cuántas hamburguesas debe encargar para el partido que se jugará el sábado siguiente. La compañía quiere basar su encargo en el número de entradas vendidas por anticipado hasta el lunes en cuestión. Los siguientes datos muestran las ventas anticipadas de entradas y el número de hamburguesas vendidas en los anteriores partidos del año. Todos los datos están dados en miles.

Ventas de entradas hasta el lunes	Hamburguesas vendidas
29,4	19,5
21,4	16,2
18,0	15,3
25,2	18,0
32,5	20,4
23,9	16,8

Si un lunes estuvieran vendidas por anticipado 26 000 entradas del partido del próximo sábado, determine el intervalo de predicción, al 95% de confianza, para la cantidad de hamburguesas que se venderán durante el partido.

Solución Si se ejecuta el programa 12-1 se obtiene la siguiente salida cuando se requiere la predicción de la respuesta futura y si el valor de la entrada es 26.

La respuesta predicha es 18,04578

W = 0,3381453

Puesto que $t_{4,\,0{,}025} = 2{,}776$, se ve, de la salida del programa, que el intervalo de predicción al 95% de confianza es

$$18{,}046 \pm 2{,}776(0{,}338) = 18{,}046 \pm 0{,}938$$

Es decir, con una confianza del 95%, el número de hamburguesas que se venderán estará comprendido entre 17 108 y 18 984 unidades. ■

Problemas

1. Utilice los siguientes datos para:

   (a) Predecir la respuesta al valor de entrada $x = 4$.

   (b) Determinar un intervalo que, con una confianza del 95%, contenga a la respuesta del apartado (a).

x	*y*
1	5
2	8
5	15

2. Un directivo de una gran fábrica de automóviles pretende estudiar la relación existente entre la edad de un trabajador y su nivel de absentismo. Se recogieron los siguientes datos relativos a 10 trabajadores elegidos aleatoriamente.

Edad	40	28	34	27	21	38	19	55	31	35
Días de ausencia	1	6	6	9	12	4	13	2	5	3

   (a) Haga una predicción del número de días de absentismo para un trabajador de 42 años de edad.

   (b) Determine, al 95% de confianza, un intervalo de predicción para el valor pedido en el apartado (a).

3. Los siguientes datos fueron registrados recientemente por un economista que quería conocer la relación existente entre las rentas de determinadas familias y las proporciones de renta gastadas en alimentación. Cada familia en cuestión estaba compuesta por el matrimonio y dos hijos de entre diez y veinte años.

Renta (en miles de dólares)	**Porcentaje gastado en alimentación**
14	35
18	33
22	32
28	29
35	23
39	19
42	17

(a) Encuentre la recta de regresión estimada.

(b) Haga una predicción de la cantidad que gasta en comida una familia de 4 miembros que gane 31 000 dólares.

(c) Determine un intervalo de predicción, al 95% de confianza, para el valor pedido en (b).

(d) Repita el apartado (c), pero, en esta ocasión, obtenga un intervalo de predicción al 99% de confianza.

4. Los siguientes datos relacionan las calificaciones obtenidas por 10 estudiantes en el examen de acceso a la universidad con sus calificaciones medias obtenidas en el primer curso universitario.

Calificación en el examen de entrada	Puntuación media del primer curso
88	3,2
74	2,7
70	2,3
77	2,9
83	2,8
94	3,6
92	3,0
81	2,8
85	3,3
92	3,1

(a) Haga una predicción de la calificación media en el primer curso de universidad de un estudiante, no incluido en los datos, que hubiera obtenido una calificación de 88 puntos en el examen de acceso.

(b) Obtenga un intervalo de predicción, al 90% de confianza, para el valor pedido en el apartado (a).

(c) Contraste la hipótesis, al nivel de significación del 5%, de que la puntuación media en el primer curso universitario es independiente de su calificación en el examen de acceso.

5. El vidrio juega un papel importante en las investigaciones criminales, porque en la actividad criminal a menudo se producen roturas de ventanas u otros objetos de vidrio. Como los fragmentos de este material se conservan fácilmente en las ropas de los criminales, es importante poder identificar que dichos fragmentos provienen de la escena del crimen. Dos propiedades físicas del cristal que resultan útiles para esta labor de identificación son su índice de refracción, que es relativamente fácil de medir, y su densidad, cuya medición es mucho más compleja. Sin embargo, la medición de densidades resulta más fácil cuando se tiene un buen estimador de su valor antes de que se fije el tipo de experimento de laboratorio necesario para su determinación exacta. En consecuencia, sería muy útil que se pudiera utilizar el índice de refracción de un fragmento de cristal para estimar su densidad.

Los siguientes datos relacionan el índice de refracción con la densidad de 12 trozos de cristal.

Índice de refracción	Densidad	Índice de refracción	Densidad
1,514	2,480	1,516	2,484
1,515	2,482	1,517	2,486
1,516	2,480	1,518	2,495
1,517	2,490	1,519	2,498
1,517	2,482	1,522	2,511
1,520	2,505	1,525	2,520

(a) Haga una predicción de la densidad de un fragmento de cristal cuyo índice de refracción sea de 1,520.

(b) Determine un intervalo que, con una confianza del 95%, contenga la densidad de un fragmento de vidrio cuyo índice de refracción sea de 1,520.

6. Los siguientes datos clasificatorios se refieren a las edades de pubertad de 20 pares de madres-hijas. Los datos x se refieren a la edad de pubertad de las madres, mientras que los datos Y se refieren a la de las hijas.

$$\bar{x} = 12{,}8 \qquad \bar{Y} = 12{,}9$$

$$S_{xx} = 36{,}5 \qquad S_{YY} = 42{,}4 \qquad S_{xY} = 24{,}4$$

(a) Encuentre la recta de regresión estimada.

(b) Utilice la fórmula computacional dada al final de la sección 12.4 para computar SS_R.

(c) Al nivel de significación del 5%, contraste la hipótesis de que $\beta = 0$.

(d) Si una madre alcanzó la pubertad a la edad de 13,8 años, determine un intervalo que, con una confianza del 95%, contenga la edad de pubertad de su hija.

7. Los siguientes datos relacionan las calificaciones medias en cursos de contabilidad con los salarios iniciales de ocho recién graduados en contabilidad.

Calificación media	Salario inicial (en miles de dólares)
3,4	42
2,5	29
3,0	33
2,8	32
3,7	40
3,5	44
2,7	30
3,1	35

(a) Haga una predicción del salario anual de un recién graduado cuya calificación media en los cursos de contabilidad fue de 2,9.

(b) Determine un intervalo que contenga el salario anual del apartado (a) con una confianza del 95%.

(c) Repita los apartados (a) y (b) para un graduado que haya obtenido 3,6 como calificación media de los cursos de contabilidad.

8. Los siguientes datos muestran los precios medios de los libros referenciados en la revista *Science* desde 1987 a 1992.

Año	Precio (en dólares)
1987	47,37
1988	54,05
1989	54,58
1990	54,43
1991	54,08
1992	57,58

Obtenga un intervalo que contenga el precio medio de los libros referenciados por *Science* en 1993, con un 95% de confianza.

12.8 Coeficiente de determinación

Supongamos que pretendemos medir la variación del conjunto de valores de respuesta $Y_1, \ldots, Y_n$ correspondientes al conjunto de valores de entrada $x_1, \ldots, x_n$. Una medida estadística estándar de la variación del conjunto de valores $Y_1, \ldots, Y_n$ viene dada por

$$S_{YY} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

Por ejemplo, si todos los Y_i son iguales –y, por consiguiente, iguales a $\overline{Y}$–, SS_{YY} será igual a 0.

La variación implícita en los valores Y_i se deriva de dos factores. En primer lugar, como los valores de entrada x_i son distintos, las variables de respuesta Y_i tienen medias diferentes, lo cual produce cierta variación en sus valores. En segundo lugar, la variación también proviene del hecho de que, aunque se tengan en cuenta las diferencias entre los valores de entrada, cada una de las variables de respuesta Y_i tiene una varianza σ^2 y, por consiguiente, no coincidirán exactamente con el valor predicho para su entrada x_i.

Consideremos ahora la cuestión de qué parte de la variación de los valores de la variable de respuesta se debe a los diferentes valores de entrada, y de qué parte se debe a la

varianza inherente a las respuestas aún cuando se hayan tenido en cuenta los valores de entrada. Para contestar a esta pregunta, observe que

$$\mathrm{SS}_R = \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} x_i)^2$$

mide el resto de variación en los valores de respuesta después de haber tenido en cuenta los distintos valores de entrada. Así pues,

$$S_{YY} - \mathrm{SS}_R$$

representa la parte de variación en las variables de respuesta que viene *explicada* por los diferentes valores de entrada; en consecuencia, la expresión R^2 definida por

$$R^2 = \frac{S_{YY} - \mathrm{SS}_R}{S_{YY}} = 1 - \frac{\mathrm{SS}_R}{S_{YY}}$$

representa la proporción de variación en las variables de respuesta que es *explicada* por los distintos valores de entrada.

Definición

R^2 recibe el nombre de *coeficiente de determinación.*

El coeficiente de determinación R^2 tiene un valor comprendido entre 0 y 1. Un valor de R^2 próximo a 1 indica que la mayor parte de la variación de la variable de respuesta viene *explicada* por los distintos valores de entrada, mientras que un valor de R^2 próximo a 0 indica que muy poca variación es *explicada* por los diferentes valores de entrada.

Ejemplo 12.9 En el ejemplo 12.5, en el que se relaciona la altura de un hijo con la de su padre, la salida de Programa 12-1 muestra que

$$S_{YY} = 38{,}521 \qquad \mathrm{SS}_R = 1{,}497$$

Por consiguiente,

$$R^2 = 1 - \frac{1{,}497}{38{,}531} = 0{,}961$$

En otras palabras, el 96% de la variación de las alturas de los 10 individuos se *explica* a partir de las alturas de sus padres. El restante 4% (inexplicado) de la variación se debe a la varianza de la altura de un hijo después de que la altura del padre se haya tenido en cuenta. (Esto es, se debe a σ^2, la varianza de la variable aleatoria de error.) ■

El valor de R^2 se utiliza a menudo como indicador de la bondad de ajuste de los datos al modelo de regresión. Un valor de R^2 próximo a 1 indica un buen ajuste, mientras que un valor próximo a 0 indica que el ajuste es pobre. En otras palabras, si el modelo de regresión es capaz de explicar la mayor parte de la variación de la variable de respuesta, se debe considerar que se ajusta bien a los datos.

Ejemplo 12.10 En el ejemplo 12.8, donde se relaciona el número de hamburguesas vendidas en un partido de fútbol con la venta adelantada de entradas del partido, el Programa 12-1 computa

$$S_{YY} = 19{,}440 \qquad \text{SS}_R = 0{,}390$$

Por consiguiente,

$$R^2 = 1 - \frac{0{,}390}{19{,}440} = 0{,}98$$

y, por tanto, el 98% de la variación del número de hamburguesas vendidas en los seis partidos se explica por la venta adelantada de entradas para dichos partidos. (Grosso modo, un 98% de las ventas se *explica* por la venta adelantada de entradas en dichos partidos.) ■

Problemas

1. Una empresa inmobiliaria recogió la siguiente información relativa a los precios de venta de casas con tres dormitorios en un determinado barrio y a los tamaños de dichas casas. (Las superficies habitables vienen dadas en unidades de 1000 pies cuadrados, mientras que los precios de venta están en unidades de 1000 dólares)

Superficie habitable	Precio de venta
2,3	240
1,8	212
2,6	253
3,0	280
2,4	248
2,3	232
2,7	260

(a) Dibuje los datos en un diagrama de dispersión.

(b) Determine la recta de regresión estimada.

(c) ¿Qué porcentaje del precio de venta viene explicado por la superficie habitable?

(d) Una casa con un tamaño de 2000 pies cuadrados acaba de salir al mercado. Determine un intervalo que incluya el precio de venta de esa casa, con una confianza del 95%.

2. Determine el R^2 para el siguiente conjunto de datos:

x	y
2	10
3	16
5	22

3. Medir directamente la cantidad de proteínas en una muestra de hígado resulta difícil y exige mucho tiempo. Por ello, los laboratorios médicos habitualmente utilizan el hecho de que la cantidad de proteína está muy relacionada con la cantidad de luz que puede absorber la muestra. Como consecuencia, un espectrómetro que emite luz se enfoca hacia una solución que contiene la muestra de hígado, y la cantidad de luz absorbida se utiliza para estimar la cantidad de proteínas.

   Este procedimiento se probó sobre cinco muestras de hígado cuyas cantidades de proteínas eran conocidas, con los siguientes datos resultantes:

Luz absorbida	**Cantidad de proteínas (en mg)**
0,44	2
0,82	16
1,20	30
1,61	46
1,83	55

   (a) Calcule el coeficiente de determinación.

   (b) ¿Este método parece ser una forma razonable de estimar la cantidad de proteínas en una muestra de hígado?

   (c) Si la luz absorbida ha sido de 1,5, ¿cuál es el estimador de la cantidad de proteínas?

   (d) Determine un intervalo de predicción para el valor del apartado (c) que tenga una confianza del 90%.

4. Obtenga el coeficiente de determinación para los datos del problema 1 de la sección 12.2.

5. Calcule el coeficiente de determinación para los datos del ejemplo 12.6.

6. A un comerciante de coches nuevos le interesa encontrar la relación existente entre el número de vendedores que trabajan durante el fin de semana y el número de coches vendidos. Los datos siguientes se recogieron durante seis domingos consecutivos:

Número de vendedores	Número de coches vendidos
5	22
7	20
4	15
2	9
4	17
8	25

(a) Determine la recta de regresión estimada.

(b) ¿Cuál es el coeficiente de determinación?

(c) ¿Qué parte de variación en el número de coches vendidos se explica por el número de vendedores?

(d) Contraste la hipótesis nula de que el número medio de ventas no depende del número de vendedores.

7. Encuentre el coeficiente de determinación en el problema 8 de la sección 12.4.

12.9 Coeficiente de correlación muestral

Consideremos un conjunto de pares de datos (x_i, Y_i), $i = 1, \ldots, n$. En la sección 3.7 se definió el *coeficiente de correlación muestral* para este conjunto de datos como

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Allí se hizo notar que r proporcionaba una medida del grado en el que los valores altos de x se corresponden con los valores altos de Y y en el que los valores bajos de x se corresponden con los valores bajos de Y. Un valor de r próximo a $+1$ indica que los valores altos de x se encuentran fuertemente asociados con los valores altos de Y y que los valores pequeños de x están fuertemente asociados con los valores pequeños de Y, mientras que un valor de r próximo a -1 indica que los valores altos de x están fuertemente asociados con los valores bajos de Y y que los valores pequeños de x están fuertemente asociados con los valores grandes de Y.

Con la notación de este capítulo, r puede expresarse como

$$r = \frac{S_{xY}}{\sqrt{S_{xx}S_{YY}}}$$

Si se utiliza la identidad

$$\mathrm{SS}_R = \frac{S_{xx}S_{YY} - S_{xY}^2}{S_{xx}}$$

que fue indicada al final de la sección 12.4, se puede demostrar que el valor absoluto del coeficiente de correlación muestral r se puede expresar como

$$|r| = \sqrt{1 - \frac{\mathrm{SS}_R}{S_{YY}}}$$

Esto es,

$$|r| = \sqrt{R^2}$$

y, por consiguiente, excepto el signo que indica si es positivo o negativo, el coeficiente de correlación muestral es igual a la raíz cuadrada del coeficiente de determinación. El signo de r coincide con el de $\hat{\beta}$.

Todo esto da un sentido adicional al coeficiente de correlación muestral. Por ejemplo, si el conjunto de datos tiene un coeficiente de correlación muestral r igual a 0,9, esto implica que el modelo de regresión lineal simple para estos datos explica un 81% (puesto que $R^2 = 0{,}9^2 = 0{,}81$) de la variación de los valores de la respuesta. Es decir, el 81% de la variación de los valores de la respuesta se explica por los diferentes valores de entrada.

Problemas

1. Calcule los coeficientes de determinación y de correlación muestral para los siguientes conjuntos de datos apareados.

(a)

x	y
2	4
3	5
5	9

(b)

x	y
4	2
5	3
9	5

¿Qué se puede concluir?

*2. Demuestre que el coeficiente de correlación muestral de un conjunto dado de pares de datos (u_i, v_i) es el mismo con independencia de que los u_i sean considerados como valores de entrada o como valores de respuesta.

3. Encuentre el coeficiente de correlación muestral si el coeficiente de determinación y la recta de regresión estimada son:

(a) $R^2 = 0{,}64, \quad y = 2x + 4$

(b) $R^2 = 0{,}64, \quad y = 2x - 4$

(c) $R^2 = 0{,}64, \quad y = -2x + 0{,}4$

(d) $R^2 = 0{,}64, \quad y = -2x - 0{,}4$

4. Si el coeficiente de correlación muestral es 0,95, ¿qué parte de la variación de las respuestas es explicada por los distintos valores de entrada?

5. Los siguientes datos muestran las edades de un grupo de esposas y sus maridos en el momento de la boda. Antes de observar los datos, ¿qué signo (positivo, negativo o próximo a cero) se puede esperar que tenga el coeficiente de correlación muestral?

Edad de la esposa	18	24	40	33	30	25
Edad del marido	21	29	51	30	36	25

   (a) Suponiendo que la edad de la esposa sea la entrada, encuentre la recta de regresión estimada para determinar la edad del marido.

   (b) Suponiendo que la edad del marido sea la entrada, encuentre la recta de regresión estimada para determinar la edad de la esposa.

   (c) Obtenga los coeficientes de determinación y de correlación muestral en la situación descrita en el apartado (a).

   (d) Obtenga los coeficientes de determinación y de correlación muestral en la situación descrita en el apartado (b).

6. Calcule el coeficiente de correlación muestral para el problema 6 de la sección 12.7.

12.10 Análisis de los residuos: evaluación del modelo

El paso inicial para comprobar si el modelo de regresión lineal simple

$$Y = \alpha + \beta x + e$$

donde e es una variable aleatoria normal con media 0 y varianza 1, es apropiado en una situación dada consiste en analizar el diagrama de dispersión. Realmente, a menudo esto resulta suficiente para que uno se convenza de que el modelo de regresión lineal simple es, o no es, correcto. Cuando el diagrama de dispersión no sea por sí mismo suficiente para corroborar el modelo, primero se deberán calcular los estimadores de mínimos cuadrados A y B, para analizar después los residuos $Y_i - (A + Bx_i)$. El análisis comienza con una normalización, o estandarización, de los residuos que se dividirán entre $\sqrt{\text{SS}_R/(n-2)}$, el estimador de la desviación típica de los Y_i. Los valores resultantes

$$\frac{Y_i - (A + Bx_i)}{\sqrt{\text{SS}_R/(n-2)}}, \qquad i = 1, \ldots, n$$

se denominan *residuos estandarizados.*

Cuando el modelo de regresión lineal simple es correcto, los residuos estandarizados son aproximadamente variables aleatorias normales estándar e independientes y, por tanto, deberían estar distribuidos aleatoriamente alrededor de 0, con aproximadamente un 95% de sus valores comprendidos entre -2 y 2 (puesto que $P\{-1{,}96 < Z < 1{,}96\} = 0{,}95$). Adicionalmente, un gráfico de los residuos estandarizados no debería reflejar ningún patrón específico de conducta. De hecho, la existencia de algún patrón determinado en el gráfico debería hacernos ser cautos acerca de la validez del modelo asumido de regresión lineal simple.

La figura 12.9 muestra tres diagramas de dispersión distintos junto con sus residuos estandarizados asociados. En el primero de ellos, como indican tanto el diagrama de dispersión como la naturaleza aleatoria de los residuos estandarizados, parece que los datos se ajustan bastante bien al modelo lineal. El segundo gráfico de los residuos muestra un patrón de conducta reconocible, en el sentido de que aparentemente los residuos primero decrecen y luego crecen a medida que los niveles de entrada crecen. A menudo, esto significa que se necesitan términos de mayor orden (superior al lineal) para describir la relación existente entre las entradas y las salidas. En realidad, eso se puede reconocer también en el diagrama de dispersión de este caso. El tercer gráfico de residuos estandarizados muestra igualmente un patrón de conducta, en el sentido de que los valores absolutos de los residuos, al igual que sus cuadrados, aparentemente crecen a medida que aumentan las entradas. Esto suele indicar que las varianzas de las respuestas no se mantienen constantes sino que, por el contrario, aumentan cuando así lo hacen las entradas.

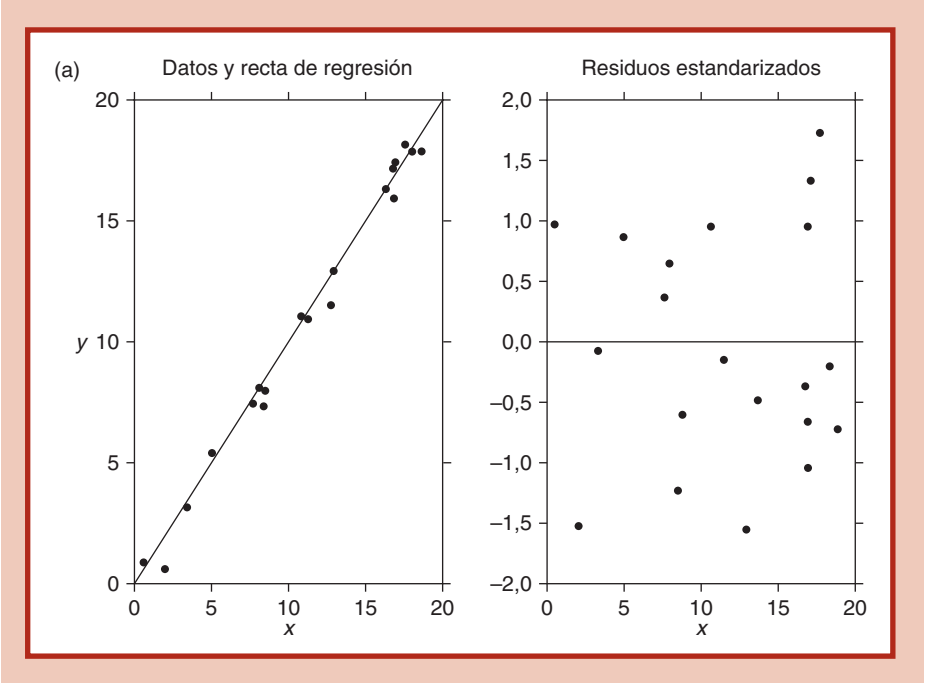
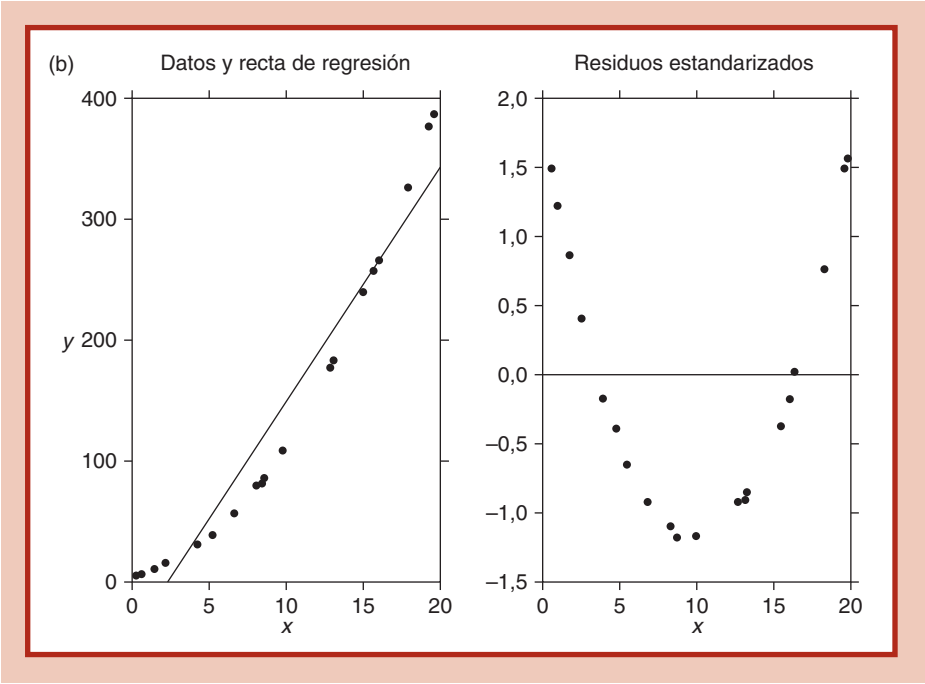


Figura 12.9 (*Continúa*)



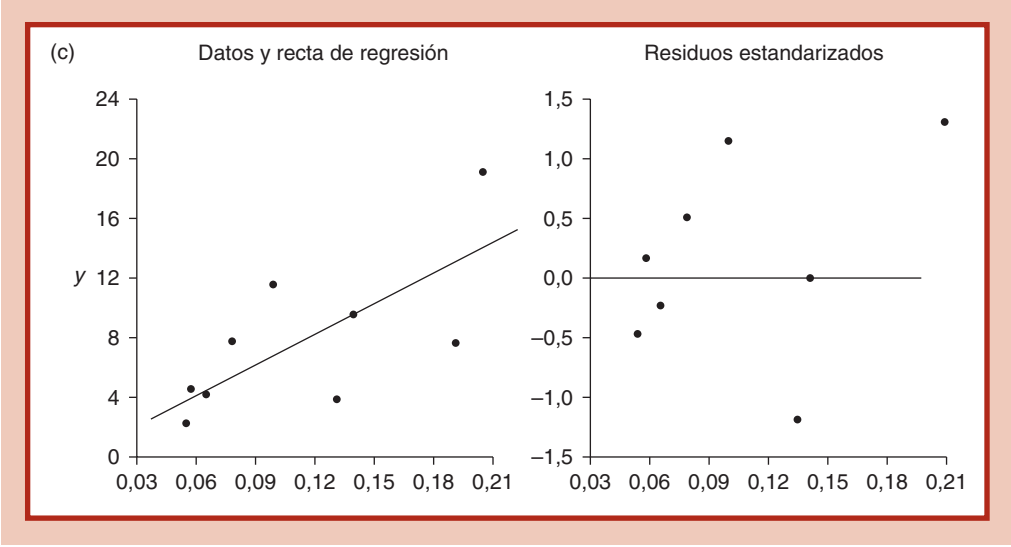


Figura 12.9 Tres diagramas de dispersión y sus residuos estandarizados asociados.

Problemas

1. Represente gráficamente los residuos estandarizados utilizando los datos del ejemplo 1 de la sección 12.3. ¿Qué conclusiones se pueden extraer sobre la validez del modelo de regresión lineal simple asumido?
2. Represente gráficamente los residuos estandarizados utilizando los datos del ejemplo 6 de la sección 12.3. ¿Qué conclusiones se pueden extraer sobre la validez del modelo de regresión lineal simple asumido?

12.11 Modelo de regresión lineal múltiple

Hasta ahora nos ha interesado predecir el valor de una respuesta a partir del valor de una única variable de entrada. Sin embargo, en muchas ocasiones la respuesta depende de diferentes variables de entrada.

Ejemplo 12.11 En experimentos de laboratorio, dos factores que a menudo afectan al porcentaje de producto obtenido son la temperatura y la presión a las que se llevó a cabo el experimento. Los siguientes datos detallan los resultados de cuatro experimentos independientes. Para cada experimento se muestran la temperatura (en grados Fahrenheit), la presión (en libras por pulgada cuadrada) a la que se desarrolló el experimento y el porcentaje de producto obtenido.

Experimento	Temperatura	Presión	Porcentaje de producto
1	140	210	68
2	150	220	82
3	160	210	74
4	130	230	80

Supongamos que estamos interesados en hacer una predicción del valor de respuesta Y a partir de los valores de k variables de entrada $x_1, x_2, \ldots, x_k$.

Definición

El modelo de *regresión lineal múltiple* asume que la respuesta Y depende de los valores de entrada x_i, $i = 1, \ldots, k$, a través de la relación

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e$$

En esta expresión, $\beta_0, \beta_1, \ldots, \beta_k$ son los parámetros de regresión y e es una variable aleatoria de error con media 0. Los parámetros de regresión no serán inicialmente conocidos y se tendrán que estimar a partir de un conjunto de datos.

Colección	Entrada 1	Entrada 2	...	Entrada k	Respuesta
1	x_{11}	x_{12}	...	x_{1k}	y_1
2	x_{21}	x_{22}	...	x_{2k}	y_2
3	x_{31}	x_{32}	...	x_{3k}	y_3
⋮					
n	x_{n1}	x_{n2}	...	x_{nk}	y_n

Figura 12.10 Datos de n experimentos.

Supongamos que se dispone de un conjunto de n respuestas correspondientes a n colecciones diferentes de k valores de entrada. Para $i = 1, \ldots, n$, denotemos por y_i la respuesta i-ésima y sean $x_{i1}, x_{i2}, \ldots, x_{ik}$ los k valores de entrada correspondientes a dicha respuesta. Así, por ejemplo, y_1 fue la respuesta obtenida cuando los valores de entrada fueron $x_{11}, x_{12}, \ldots, x_{1k}$. El conjunto completo de datos se muestra en la figura 12.10.

Ejemplo 12.12 En el ejemplo 12.11 existen dos variables de entrada, la temperatura y la presión; por tanto, $k = 2$. Existen cuatro resultados experimentales, de donde $n = 4$. Los valores x_{i1} se refieren a la temperatura y los de x_{i2} a la presión, ambos del experimento i. El valor y_i es el porcentaje de producto (respuesta) del experimento i. Así, por ejemplo,

$$x_{31} = 160 \qquad x_{32} = 210 \qquad y_3 = 74 \quad \blacksquare$$

Para estimar de nuevo los parámetros de regresión, tal como se hizo en el caso de la regresión lineal simple, se utilizará el método de mínimos cuadrados. Esto es, se empezará observando que si $B_0, B_1, \ldots, B_k$ son los estimadores de los parámetros de regresión $\beta_0, \beta_1, \ldots, \beta_k$, el estimador de la respuesta cuando los valores de entrada son $x_{i1}, x_{i2}, \ldots, x_{ik}$ viene dado por

$$\text{Respuesta estimada} = B_0 + B_1 x_{i1} + B_2 x_{i2} + \cdots + B_k x_{ik}$$

Puesto que la respuesta observada fue y_i, se ve que la diferencia entre la respuesta observada y la que se hubiera predicho si se hubieran usado los estimadores $B_0, B_1, \ldots, B_k$ es

$$\epsilon_i = y_i - (B_0 + B_1 x_{i1} + B_2 x_{i2} + \cdots + B_k x_{ik})$$

Por consiguiente, ϵ_i puede ser considerado como el error que se habría cometido si se hubieran utilizado los estimadores B_i, $i = 1, \ldots, k$. Aquellos estimadores que minimizan la suma de los cuadrados de los errores se denominan *estimadores de (por) mínimos cuadrados*.

Los estimadores de mínimos cuadrados de los parámetros de la regresión son aquellos valores de los B_i que hacen que

$$\sum_{i=1}^{n} \epsilon_i^2$$

sea lo más pequeña posible.

Los cálculos necesarios para obtener los estimadores de mínimos cuadrados son complicados y no se van a abordar aquí. En su lugar, remitimos al Programa 12-2 para que éste haga los cómputos. La salida de este programa incluye los estimadores de los parámetros de regresión. Adicionalmente, el programa suministra los valores de respuesta predichos correspondientes a conjuntos específicos de valores de entrada. Esto es, si el usuario introduce los valores $x_1, x_2, \ldots, x_k$, el ordenador imprimirá el valor $B(0) + B(1)x_1 + \cdots + B(k)x_k$, donde $B(0), B(1), \ldots, B(k)$ representan los estimadores de mínimos cuadrados de los parámetros de regresión.

Ejemplo 12.13 Los siguientes datos relacionan las tasas de suicidio, y, con el tamaño de la población, x_1, y la tasa anual de divorcios, x_2, en ocho ciudades distintas.

Ciudad	Población (en miles)	Tasa de divorcio por 100 000 habitantes	Tasa de suicidio por 100 000 habitantes
Akron, OH	679	30,4	11,6
Anaheim, CA	1420	34,1	16,1
Buffalo, NY	1349	17,2	9,3
Austin, TX	296	26,8	9,1
Chicago, IL	3975	29,1	8,4
Columbia, SC	323	18,7	7,7
Detroit, MI	2200	32,6	11,3
Gary, IN	633	32,5	8,4

(a) Ajuste un modelo de regresión lineal múltiple a estos datos. Es decir, ajuste un modelo de la forma

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$$

donde Y es la tasa de suicidio, x_1 es el tamaño de la población y x_2 es la tasa de divorcio.

(b) Haga una predicción de la tasa de suicidio en una ciudad con 400 000 habitantes de población y con una tasa de divorcios de 28,4 divorcios por cada 100 000 habitantes.

Perspectiva histórica

Método de mínimos cuadrados

La primera publicación en donde se detalla en método de mínimos cuadrados se debe al científico matemático francés Adrien Marie Legendre, en 1805. Legendre presentó este método en el apéndice de su libro *Novelles méthodes pour la détermination des orbites des comètes (Nuevos métodos para la determinación de las órbitas de los cometas).* Tras explicar el método, Legendre desarrolló un ejemplo, usando datos procedentes de mediciones del arco meridiano francés realizadas en 1795, en el que $k = 2$ y $n = 5$. En 1809, Karl Friedrich Gauss publicó una justificación del método de mínimos cuadrados, en la que se resaltaba que la normal era la distribución del término de error. En este trabajo, Gauss comenzó una controversia, pues mantenía que él había estado usando el método desde 1795. Indicó que había utilizado el método de mínimos cuadrados en 1801 para localizar el asteroide perdido Ceres. Este asteroide, el mayor del sistema solar y el primero en ser descubierto, fue localizado el 1 de enero de 1801 por el astrónomo italiano Giuseppe Piazzi del Observatorio de Palermo. Piazzi lo observó durante 40 días consecutivos tras los cuales el asteroide, que tenía muy poca luminosidad, desapareció de la vista. Con el objetivo de que otros científicos intentaran determinar su trayectoria, Piazzi publicó los datos relativos a sus observaciones. En poco tiempo y sin ninguna explicación, Gauss publicó una predicción de la órbita del asteroide. Inmediatamente después, se encontró Ceres casi en la posición exacta predicha por Gauss.

La disputa entre Legendre y Gauss sobre la autoría del método fue muy acalorada. En la edición de 1820 de su libro, Legendre añadió un ataque a Gauss, que atribuyó al escritor anónimo Monsieur***. Gauss, por su parte, solicitó el testimonio de colegas a los que había comunicado su método con anterioridad a 1805. Hoy en día, la reclamación de Gauss, en el sentido de que él conocía el método de mínimos cuadrados con anterioridad a Legendre, es generalmente aceptada por la mayoría de los científicos actuales. (Gauss es conocido porque normalmente pasaban varios años antes de que publicara sus trabajos.) Sin embargo, muchos científicos mantienen que se debería asignar la autoría atendiendo tan sólo a la fecha de publicación y, en consecuencia, el mérito del descubrimiento del método de mínimos cuadrados le correspondería por derecho a Legendre.

Solución Con el Programa 12-2 se obtiene la salida siguiente.

Los estimadores de los coeficientes de regresión son los siguientes:

B(0) = 3,686646
B(1) = −2,411092E−04
B(2) = ,2485504

Si los dos valores de entrada son 400 y 28,4, la predicción de la respuesta es 10,64903. Esto es, la ecuación de regresión múltiple estimada es

$$Y = 3{,}6866 - 0{,}00024x_1 + 0{,}24855x_2$$

La tasa de suicidio predicha es

$$y = 3{,}6866 - 0{,}00024 \times 400 + 0{,}24855 \times 28{,}4$$
$$= 10{,}649$$

Es decir, la predicción de la tasa anual de suicidio para la ciudad en cuestión es de 10,649 suicidios por cada 100 000 habitantes. Puesto que el tamaño de la población es de 400 000 habitantes, lo anterior equivale a predecir que ocurrirán 42,596 suicidios al año en la dicha ciudad. ■

Problemas

1. Los datos siguientes relacionan los precios de venta, y, con la superficie habitable, x_1, el tamaño de la parcela, x_2, y el número de cuartos de baño, x_3, para 10 casas vendidas recientemente en un barrio residencial.

Precio de venta (miles de dólares)	Superficie habitable (pies cuadrados)	Tamaño de la parcela (acres)	Número de cuartos de baño
170	1300	0,25	1
177	1450	0,30	1,5
191	1600	0,30	2
194	1850	0,45	2
202	2100	0,40	2
210	2000	0,40	2,5
214	2100	0,50	2
228	2400	0,50	2,5
240	2700	0,50	2,5
252	2600	0,70	3

   (a) Ajuste un modelo de regresión lineal múltiple a los datos anteriores.

   (b) Haga una predicción del precio de una casa con 2500 pies cuadrados de superficie habitable, 0,4 acres de parcela y 2 cuartos de baño.

   (c) Si la casa del apartado (b) tuviera tres cuartos de baño, ¿cuál sería su precio?

2. En el ejemplo 12.11 haga una predicción del porcentaje de producto que se obtendría si el experimento se llevara a cabo a una temperatura de 150 grados Fahrenheit y a una presión de 215 libras por pulgada cuadrada.

3. Ajuste un modelo de regresión lineal múltiple al siguiente conjunto de datos:

x_1	x_2	x_3	x_4	y
1	3	5	9	121
2	4	4	10	165
1,5	8	2	14	150
3	9	3	8	170
1	11	4	12	140

Prediga el valor de la respuesta correspondiente a los valores de entrada

$$x_1 = 2 \qquad x_2 = 7 \qquad x_3 = 3 \qquad x_4 = 13$$

4. El siguiente conjunto de datos se refiere a los transplantes de corazón de Stanford. Relaciona los tiempos de supervivencia de los pacientes de trasplante de corazón, con sus edades en el momento del trasplante y sus puntuaciones en una prueba de incompatibilidad, que es un indicador de la aceptabilidad del corazón por parte del receptor.

Tiempo de supervivencia (en días)	Puntuación de incompatibilidad	Edad
624	1,32	51,0
1350	0,87	54,1
64	1,89	54,6
46	0,61	42,5
1024	1,13	43,4
280	1,12	49,5
10	2,76	55,3
60	0,69	64,5
836	1,58	45,0
136	1,62	52,0
730	0,96	58,4
39	1,38	42,8

(a) Ajuste un modelo de regresión lineal múltiple a estos datos.

(b) Estime el tiempo de supervivencia de un paciente al que se le trasplanta un corazón a los 50 años de edad y cuya puntuación de incompatibilidad haya resultado ser de 1,46.

5. En una compañía siderúrgica se van a producir planchas de acero que contienen un 0,15% de cobre y que han sido fabricadas a una temperatura de fundición de 1150 grados Fahrenheit. La compañía está interesada en estimar la dureza media del este acero, para lo cual ha recogido los datos siguientes relativos a 10 tipos diferentes de planchas de acero producidas a distintas temperaturas de fundición y con distintos contenidos de cobre.

Dureza	Contenido de cobre	Temperatura de fundición
79,2	0,02	1050
64,0	0,03	1200
55,7	0,03	1250
56,3	0,04	1300
58,6	0,10	1300
49,8	0,09	1450
51,1	0,12	1400
61,0	0,09	1200
70,4	0,15	1100
84,3	0,16	1000

Estime la dureza media del acero que se va a producir.

Términos clave

Regresión lineal simple: Modelo que relaciona una variable de respuesta Y con una variable de entrada x por medio de la ecuación

$$Y = \alpha + \beta x + e$$

En ésta, α y β son los parámetros del modelo de regresión, y e es la variable aleatoria de error.

Variable dependiente: Otro nombre que recibe la variable de respuesta.

Variable independiente: Otro nombre que recibe la variable de entrada.

Método de mínimos cuadrados: Método para obtener los estimadores de los parámetros de regresión α y β. Toma como estimadores aquellos valores que minimizan la suma de cuadrados de las diferencias entre las respuestas observadas y las predichas.

Regresión a la media: Este fenómeno ocurre cuando el parámetro de regresión β está estrictamente comprendido entre 0 y 1. Produce que la respuesta media correspondiente a un valor de entrada x sea mayor que x cuando x es pequeño, y que sea menor que x cuando x es grande. Este fenómeno es habitual en situaciones de contrastación-recontrastación.

Falacia de la regresión: Creencia, en situaciones de contrastación-recontrastación, de que el fenómeno de la regresión a la media tiene una causa justificativa, cuando en realidad es simplemente un subproducto de las fluctuaciones aleatorias.

Coeficiente de determinación: Estadístico cuyo valor indica la proporción de la variación de los valores de respuesta que se debe a los diferentes valores de entrada.

Coeficiente de correlación muestral: Su valor absoluto coincide con la raíz cuadrada del coeficiente de determinación. Su signo es igual al del estimador del parámetro de regresión β.

Regresión lineal múltiple: Modelo que relaciona una variable de respuesta Y con una colección de k variables de entrada $x_1, x_2, \ldots, x_k$ por medio de la ecuación

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e$$

Resumen

El modelo de regresión lineal simple relaciona el valor de una variable aleatoria *de respuesta Y* con el valor de una variable *de entrada x* mediante la ecuación

$$Y = \alpha + \beta x + e$$

Los parámetros α y β son los *parámetros de regresión* que se deben estimar a partir de los datos. El término e es una variable aleatoria *de error* que tiene una media 0.

Se utiliza el *método de mínimos cuadrados* para estimar los parámetros de regresión α y β. Supongamos que se han llevado a cabo varios experimentos en los que se han fijado previamente los niveles de las entradas x_i, $i = 1, \ldots, n$. Denotemos por Y_i, $i = 1, \ldots, n$ sus correspondientes respuestas. El método de mínimos cuadrados consiste en elegir como estimadores de α y β a aquellos valores de A y B que minimizan

$$\sum_{i=1}^{n} (Y_i - A - Bx_i)^2$$

Los valores de A y B que consiguen esto –llamémosles $\hat{\alpha}$ y $\hat{\beta}$– vienen dados por

$$\hat{\beta} = \frac{S_{xY}}{S_{xx}}$$

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{x}$$

donde $\overline{x}$ y $\overline{Y}$ son los valores medios de las x_i e Y_i, respectivamente, y

$$S_{xY} = \sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y}) = \sum_{i=1}^{n} x_i Y_i - n\overline{x}\overline{Y}$$

$$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2$$

La relación lineal

$$y = \hat{\alpha} + \hat{\beta}x$$

se denomina *recta de regresión estimada*.

Se asume que la variable aleatoria de error e sigue una distribución normal con media 0 y varianza σ^2, donde σ^2 es desconocida y se debe estimar a partir de los datos. El estimador de σ^2 es

$$\frac{\mathrm{SS}_R}{n - 2}$$

donde SS_R, conocido como *suma de cuadrados de los residuos*, se define por

$$\mathrm{SS}_R = \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta}x_i)^2$$

Las diferencias $Y_i - \hat{\alpha} - \hat{\beta}x_i$, que representan las discrepancias entre las respuestas observadas y sus predicciones bajo los estimadores de mínimos cuadrados, se denominan *residuos*.

La fórmula siguiente es útil para calcular SS_R con una calculadora de mano.

$$SS_R = \frac{S_{xx}S_{YY} - S_{xY}^2}{S_{YY}}$$

donde

$$S_{YY} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

Si el parámetro de regresión β es igual a 0, el valor de una respuesta no se verá afectado por su valor de entrada x. Para comprobar si esta hipótesis es plausible, se puede contrastar

$$H_0: \beta = 0 \qquad \text{frente a} \qquad H_1: \beta \neq 0$$

El contraste a nivel de significación γ se basa en el estadístico

$$TS = \sqrt{\frac{(n-2)S_{xx}}{SS_R}}\,\hat{\beta}$$

y consiste en

Rechazar H_0	si $\lvert TS\rvert \geq t_{n-2,\,\gamma/2}$
No rechazar H_0	en otro caso

Equivalentemente, si el valor de TS, es v el p valor viene dado por

$$p \text{ valor} = 2P\{T_{n-2} \geq |v|\}$$

donde T_{n-2} es una variable aleatoria t con $n - 2$ grados de libertad.

Se dice que ocurre el fenómeno de *regresión a la media* cuando el parámetro de regresión β varía entre 0 y 1. Cuando es así, la respuesta esperada correspondiente al valor de entrada x será mayor que x cuando x es pequeño, y será menor que x cuando x es grande.

El fenómeno de regresión a la media a menudo se puede ver en situaciones de contrastación-recontrastación referidas a una población homogénea. Esto se debe a que algunos de los elementos observados se comportan significativamente mejor o peor que lo que es normal simplemente debido al azar. En observaciones repetidas, por lo general se obtienen resultados menos anómalos. Es decir, aquellos valores altos en la primera observación, de alguna forma, se verían disminuidos en la segunda; mientras que los valores bajos en la primera observación tenderían a mejorar en la segunda. La creencia de que algo significativo ha propiciado la regresión a la media (por ejemplo, que los estudiantes con calificaciones menores en un primer examen se han esforzado mucho más en un segundo examen, mientras que los que consiguieron calificaciones altas en el primero han sido más perezosos en el segundo examen), cuando en realidad se debe simplemente a fluctuaciones aleatorias alrededor de la media, se conoce con el nombre de *falacia de la regresión.*

Los pares de datos de entrada-respuesta (x_i, y_i), $i = 1, \ldots, n$, se pueden utilizar para obtener un *intervalo de predicción* que contenga, con un nivel de confianza dado, una futura respuesta ante un valor de entrada x_0. En concreto, se puede asegurar, con una confianza del $100(1 - \gamma)\%$, que la respuesta al valor de entrada x_0 caerá dentro del intervalo

$$\hat{\alpha} + \hat{\beta}x_0 \pm t_{n-2,\gamma/2}W$$

donde

$$W = \sqrt{\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]\frac{\text{SS}_R}{n-2}}$$

Los valores de $\hat{\alpha}, \hat{\beta}, \bar{x}, S_{xx}$ y W se basan en los pares de datos (x_i, y_i), $i = 1, \ldots, n$, y se pueden obtener ejecutando el Programa 12-1.

El estadístico R^2 definido por

$$R^2 = 1 - \frac{\text{SS}_R}{S_{YY}}$$

se denomina *coeficiente de determinación.* Su valor, que siempre se encuentra entre 0 y 1, se puede interpretar como la proporción de variación en los valores de respuesta que viene explicada por los diferentes valores de entrada.

El estadístico r, definido por

$$r = \frac{S_{xY}}{\sqrt{S_{xx}S_{YY}}}$$

se denomina *coeficiente de correlación muestral.* Excluido el signo (que puede ser positivo o negativo), es igual a la raíz cuadrada del coeficiente de determinación. Esto es,

$$|r| = \sqrt{R^2}$$

Los valores

$$\frac{Y_i - (A + Bx_i)}{\sqrt{\text{SS}_R/(n-2)}}, \quad i = 1, \ldots, n$$

se denominan *residuos estandarizados*. La representación gráfica de estos residuos se puede utilizar para evaluar la precisión del modelo de regresión lineal.

La *regresión lineal múltiple* relaciona una variable aleatoria de respuesta Y con una colección de variables de entrada $x_1, x_2, \ldots, x_k$ de acuerdo con la ecuación

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e$$

En esta ecuación, $\beta_0, \beta_1, \ldots, \beta_k$ son los parámetros de regresión, y e es una variable aleatoria de error con media 0.

Los parámetros de regresión se estiman a partir de los datos con el método de mínimos cuadrados. Es decir, los estimadores se eligen de modo que se minimice la suma de cuadrados de las diferencias entre los valores de respuesta observados y sus valores predichos. Para obtener estos estimadores se puede usar el Programa 12-2. Este programa también calcula las predicciones de las respuestas correspondientes a una colección dada de valores de entrada.

Problemas de repaso

1. Los siguientes datos relacionan la resistencia a la rotura de ocho tejidos con el porcentaje de nailon (en vez de algodón) de estos últimos.

Porcentaje de nailon	Resistencia a la rotura (en libras)
0	160
10	240
20	325
20	340
30	395
40	450
50	510
50	520

   (a) Represente gráficamente estos datos mediante un diagrama de dispersión.

   (b) Obtenga la recta de regresión estimada.

   (c) Haga una predicción de la resistencia a la rotura de un tejido con un 50% de nailon.

   (d) Calcule un intervalo que, con un 95% de confianza, contenga el valor de la resistencia a la rotura de un tejido con un 50% de nailon.

2. Es generalmente aceptado que, si incrementa el número de unidades producidas, se puede habitualmente disminuir su coste por unidad. Los siguientes datos relacionan los costes unitarios de producción con el número de unidades producidas.

Número de unidades	10	20	50	100	150	200
Coste por unidad	9,4	9,2	9,0	8,5	8,1	7,4

   (a) Prediga el coste por unidad cuando se tiene una producción de 125 unidades.

   (b) Estime la varianza del coste indicado en el apartado (a).

   (c) Obtenga un intervalo que contenga, con un 99% de confianza, el coste por unidad cuando se tiene una producción de 110 unidades.

3. Utilice los datos relativos a las primeras 20 mujeres del conjunto de datos incluido en el Apéndice A. Supongamos que la variable de entrada es el peso y que la variable de respuesta es la presión sistólica sanguínea.

   (a) Estime los parámetros de regresión.

   (b) Obtenga un intervalo de predicción, al 95% de confianza, para la presión sistólica sanguínea de una estudiante cuyo peso sea de 120 libras.

   (c) Encuentre todas las estudiantes del Apéndice A cuyo peso esté comprendido entre 119 y 121 libras. ¿Qué porcentaje de ellas tienen presiones sanguíneas sistólicas que caen dentro del intervalo obtenido en el apartado (b)?

4. Se elije aleatoriamente a un conjunto de 10 matrimonios entre todos los pertenecientes a una comunidad dada y los 20 individuos seleccionados se someten a un test de inteligencia (IQ). Numere los matrimonios seleccionados y denotemos por x_i e y_i las puntuaciones obtenidas, respectivamente, por la esposa y el marido de la pareja i. ¿Piensa que el diagrama de dispersión de los datos resultantes podría indicar una regresión a la media? Explique por qué.

5. Varios instructores de vuelo con experiencia mantienen que habitualmente los aterrizajes aplaudidos como excepcionalmente buenos vienen seguidos de aterrizajes mucho más pobres; mientras que, por lo general, los aterrizajes criticados como muy malos vienen seguidos de aterrizajes muy mejorados. ¿Se podría concluir que las alabanzas tienden a reducir la calidad del siguiente aterrizaje, mientras que las críticas tienden a aumentar dicha calidad? ¿O existe otra explicación posible?

6. Los siguientes datos relacionan el número medio de cigarrillos fumados diariamente con el número de radicales libres encontrados en los pulmones de ocho individuos.

Número de cigarrillos	Radicales libres
0	94
10	144
14	182
5	120
18	240
20	234
30	321
40	400

   (a) Represente estos datos en un diagrama de dispersión.

   (b) Dibuje a mano una recta que se ajuste a los datos.

   (c) Determine la recta de regresión estimada y compárela con la dibujada en el apartado (b).

   (d) Haga una predicción del número de radicales libres en una persona que fuma una media de 26 cigarrillos diarios.

(e) Determine, con una confianza del 95%, un intervalo de predicción que contenga el número de radicales libres de un individuo que fume una media diaria de 26 cigarrillos.

7. Los siguientes datos muestran los precios al por menor del galón de gasolina en Estados Unidos entre los años 1990 y 2002.

Año	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002
Precio	1,16	1,14	1,13	1,11	1,11	1,15	1,23	1,23	1,06	1,17	1,51	1,46	1,36

(a) Encuentre la recta de regresión estimada.

(b) Al nivel de significación del 5%, contraste la hipótesis de que $\beta = 0$.

8. Los datos siguientes muestran las calificaciones medias de los estudiantes de secundaria que se presentaron al examen de acceso a la universidad en las materias de ciencias durante varios años, que llegan hasta 2003, excluyendo 1998.

Año	1996	1997	1999	2000	2001	2002	2003
Calificación	20,9	21,0	21,0	21,0	21,0	20,8	20,8

Fuente: Informe sobre Enseñanza Secundaria, Programa ACT, Ciudad de Iowa, IA.

(a) Haga una predicción de la calificación media de 1998.

(b) Encuentre, al 95% de confianza, un intervalo de predicción de la calificación anterior.

9. La tabla siguiente muestra el porcentaje de trabajadores afiliados a algún sindicato en los años 1984 y 1989 para una muestra de nueve Estados.

Estado	**1984**	**1989**
Alabama	27,3	23,8
Colorado	10,9	9,5
Illinois	40,9	29,8
Kentucky	27,0	21,5
Minnesota	25,7	16,4
New Jersey	25,4	24,4
Texas	15,9	13,8
Wisconsin	31,1	23,0
New York	50,4	47,2

Fuente: Estudio de Producción Grant/Thornton, Anuario

(a) El porcentaje de trabajadores de Ohio afiliados a algún sindicato fue del 41,6% en 1984. Pronostique el porcentaje correspondiente a 1989.

(b) El porcentaje de afiliación en Oklahoma fue del 17,5% en 1984. Con una confianza del 95%, construya un intervalo que contenga el porcentaje de afiliación de Oklahoma en 1989.

10. Los siguientes datos muestran las importaciones medias de petróleo (en miles de barriles por día) de Estados Unidos procedentes del Golfo Pérsico en los años comprendidos entre 1994 y 2002.

Año	1994	1995	1996	1997	1998	1999	2000	2001	2002
Importación	1728	1573	1604	1755	2136	2464	2488	2761	2269

Obtenga un intervalo que, con un 95% de confianza, contenga la importación media por día en 2002.

11. Se piensa que la resistencia a la tensión de una cierta fibra sintética depende del porcentaje de algodón y del tiempo de secado. El análisis de ocho tipos de fibra condujo a los resultados siguientes:

Porcentaje de algodón	Tiempo de secado	Resistencia a la tensión
13	2,1	212
15	2,2	221
18	2,5	230
20	2,4	219
18	3,2	245
20	3,3	238
17	4,1	243
18	4,3	242

(a) Ajuste una ecuación de regresión lineal múltiple, teniendo como variable de respuesta la resistencia a la tensión y como variables de entrada el porcentaje de algodón y el tiempo de secado.

(b) Haga una predicción de la resistencia a la tensión de una fibra sintética que tenga un porcentaje de algodón del 22% y cuyo tiempo de secado haya sido 3,5.

12. Los siguientes datos se refieren a la producción por acre de trigo en ocho plantaciones diferentes, todas ellas teniendo aproximadamente la misma calidad de tierra. Los datos relacionan la producción de trigo de cada plantación con la cantidad de lluvia caída y la cantidad de fertilizante empleado por acre.

Lluvia caída (en pulgadas)	Fertilizante (libras por acre)	Producción de trigo
15,4	100	46,6
18,2	85	45,7
17,6	95	50,4
18,4	140	66,5
24,0	150	82,1
25,2	100	63,7
30,3	120	75,8
31,0	80	58,9

(a) Estime los parámetros de regresión.

(b) Estime la producción adicional de trigo por cada pulgada adicional de lluvia.

(c) Estime la producción adicional de trigo por cada libra adicional de fertilizante.

(d) Haga una predicción de la producción de trigo en un año que haya tenido 26 pulgadas de lluvia si la cantidad de fertilizante empleado ese año fue de 130 libras por acre.

13. Un estudio completado recientemente, que hace referencia a los trabajadores municipales, intenta relacionar el grado de satisfacción en el trabajo con el sueldo y la antigüedad; para ello se seleccionó una muestra de nueve trabajadores. El grado de satisfacción fue evaluado por el propio trabajador, con una puntuación mínima de 1 y máxima de 10. Los siguientes datos muestran los resultados obtenidos.

Sueldo anual (miles de dólares)	Años en el trabajo	Satisfacción en el trabajo
47	8	5,6
42	4	6,3
54	12	6,8
48	9	6,7
56	16	7,0
59	14	7,7
53	10	7,0
62	15	8,0
66	22	7,8

(a) Estime los parámetros de regresión.

(b) ¿Qué conclusiones cualitativas se pueden extraer acerca de cómo varía el grado de satisfacción cuando el sueldo permanece constante y el número de años aumenta?

(c) Haga una predicción del grado de satisfacción de un empleado con una antigüedad de 5 años en el trabajo y con un salario anual de 51 000 dólares.

14. Suponga, en el problema 13, que el grado de satisfacción en el trabajo dependiera sólo de la antigüedad en el mismo y que, por consiguiente, se dispusiera de los siguientes datos:

Años en el trabajo	Satisfacción en el trabajo	Años en el trabajo	Satisfacción en el trabajo
8	5,6	14	7,7
4	6,3	10	7,0
12	6,8	15	8,0
9	6,7	22	7,8
16	7,0		

(a) Estime los parámetros de regresión α y β.

(b) ¿Cuál es la relación cualitativa entre los años de servicio y el grado de satisfacción? Es decir, basándose en los datos dados, ¿qué parece suceder con el grado de satisfacción a medida que los años de servicio crecen?

(c) Compare la respuesta del apartado (b) con la respuesta del apartado (b) del problema 13.

(d) ¿Qué conclusión, si es que encuentra alguna, se puede extraer de su respuesta al apartado (c)?

15. La respuesta correcta al problema 5 de la sección 12.5 consiste en rechazar la hipótesis de que no existe relación entre el consumo de cigarrillos y las tasas de cáncer de vejiga. ¿Implica este hecho que el consumo de cigarrillos conduce directamente a un mayor riesgo de contraer cáncer de vejiga o se puede pensar en otra explicación? (*Sugerencia*: ¿Existe otra variable en la que se pueda pensar que esté asociada estadísticamente tanto con el consumo de tabaco como con el cáncer de vejiga? ¿Qué tipo de recogida de datos y qué procedimiento estadístico sería recomendable para incrementar nuestro conocimiento sobre los factores que afectan a las tasas de cáncer de vejiga?)

13

Contrastes de bondad de ajuste de la chi-cuadrado

Calma, es un error discutir con los datos delante. Inconscientemente, siempre se intentará retorcerlos hasta que se ajusten a lo que cada uno mantiene.

Sherlock Holmes, *Las aventuras de Wisteria Lodge*

Pocas observaciones y mucho razonamiento conducen al error; muchas observaciones y poco razonamiento, a la verdad.

Alexis Carrel

Se considerará una población en la que cada miembro puede tomar uno cualquiera de k valores posibles. Se muestra cómo contrastar la hipótesis de que un conjunto determinado de probabilidades representa las proporciones de los miembros de la población que toman cada uno de los posibles valores distintos. Se considerarán poblaciones en las que cada miembro se clasifica de acuerdo con dos valores, y se mostrará cómo contrastar la hipótesis de que los dos valores de un miembro de la población seleccionado aleatoriamente son independientes.

13.1 Introducción

Se considera que la manipulación de los datos para conseguir que éstos corroboren una determinada hipótesis científica es un ejemplo de fraude científico. A lo largo de los años se ha cometido una gran cantidad de fraudes científicos, cuya gravedad varía desde ligeras "inconcreciones" hasta rotundas falsificaciones de los datos. Por ejemplo, uno de los casos más ilustres afectó al psicólogo educacional británico Cyril Burt. Burt fue muy reconocido en vida –de hecho, fue investido caballero de la reina de Inglaterra, convirtiéndose en Sir Cyril Burt– debido a sus investigaciones centradas en test de inteligencia de gemelos que habían crecido separadamente. Sin embargo, hoy día está generalmente aceptado que en sus trabajos publicados no solamente inventaba los datos que hacía públicos sino que también resulta cuestionable la mera veracidad de sus supuestos temas de investigación y la existencia de sus colaboradores.

Posiblemente, el más sorprendente ejemplo de fraude científico hace referencia al monje austriaco Gregor Mendel (1822-1884), al que se le considera el fundador de la moderna Teoría Genética. En 1865, Mendel publicó un artículo donde presentaba una serie de experimentos llevados a cabo con guisantes de jardín. Uno de dichos experimentos afectaba al color –amarillo o verde– de las semillas de dichos guisantes. Mendel comenzó su experimento produciendo guisantes amarillos de raza pura, una raza de guisantes en la que las plantas de cada generación tienen únicamente semillas amarillas. Después, produjo guisantes verdes de raza pura. Finalmente, Mendel cruzó guisantes amarillos puros con guisantes verdes puros. Las semillas resultantes de este cruce, conocidas como semillas *híbridas de primera generación*, resultaron ser siempre amarillas. Es decir, no resultó ninguna semilla verde en esa generación.

Tras ello, Mendel cruzó entre sí semillas de esta primera generación, para obtener semillas de segunda generación. Sorprendentemente, las semillas verdes reaparecieron aquí. De hecho, aproximadamente un 25% de las semillas de la segunda generación fueron verdes, y el 75% restante, amarillas.

En su artículo, Mendel presentó una teoría para explicar esos resultados. Según ésta, cada semilla constaba de dos entidades, en la actualidad llamadas *genes*, que conjuntamente determinan el color de la semilla. Cada gen es de uno de dos tipos posibles: tipo *a* (por amarillo) y tipo *v* (por verde). La teoría de Mendel mantenía que cada par de genes de las semillas de raza amarilla pura es siempre del tipo *a*,*a*. Esto es, los dos genes de un guisante de la raza amarilla pura son amarillos. De igual forma, el par de genes de las semillas de la raza verde pura son *v*,*v*. Mendel suponía que, cuando se cruzan dos semillas, cada descendiente resultante recibe un gen de cada progenitor. Adicionalmente, suponía también que el gen recibido de uno de sus progenitores tenía la misma probabilidad de ser uno cualquiera de su par. Así pues, cuando se cruza una semilla amarilla pura *a*,*a* con una semilla verde pura *v*,*v*, la descendencia tendrá con seguridad un gen *a* y un gen *v*; esto es, la descendencia tendrá el par de genes *a*,*v*. Puesto que todos los descendientes resultantes de cruces entre guisantes puros eran amarillos, Mendel postuló que el gen *a* era dominante frente al gen *v*, en el sentido de que una semilla con el par de genes *a*,*v* siempre mostraba el color amarillo. Véase la figura 13.1.

Veamos ahora qué sucede cuando se cruzan dos semillas de la primera generación. Observe en primer lugar que ambas semillas son híbridas, tendiendo el par de genes *a*,*v*. Observe después que, para que un descendiente sea verde, debe recibir el gen *v* de cada progenitor. Dado que es igualmente probable que cada progenitor transmita su gen *a* o su gen *v*, se tiene que la probabilidad de que ambos progenitores transmitan el gen *v* es $1/2 \times 1/2 = 1/4$.

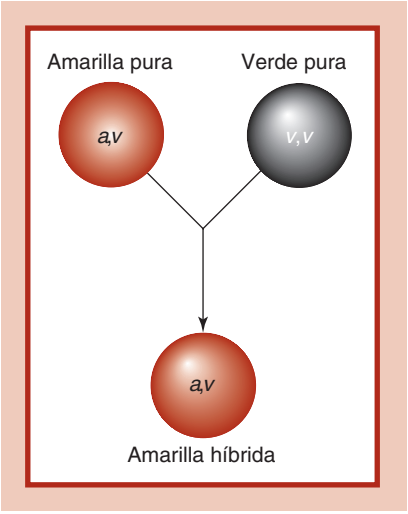


Figura 13.1 Cruce de semillas amarillas puras con semillas verdes puras.

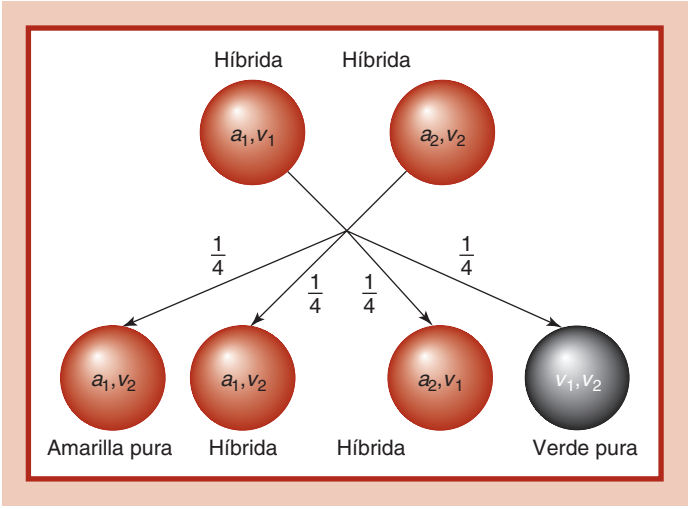


Figura 13.2 Cruce de semillas híbridas de la primera generación.

Así pues, debería resultar que, tras un gran número de cruces entre semillas de la primera generación, aproximadamente un 25% de las semillas de la siguiente generación fueran verdes. Es decir, exactamente el resultado que Mendel indicó en su artículo. Véase la figura 13.2.

El descubrimiento de Mendel fue de capital importancia para la ciencia. Pese a ello, pasó tiempo hasta que su teoría genética se convirtió en un elemento clave de la ciencia (ya que su trabajo fue ignorado durante cerca de 30 años).

No es difícil imaginar la consternación que, en 1936, produjo entre los genetistas la publicación de un artículo de R. A. Fisher en el que se analizaban los datos de Mendel y se concluía que dichos datos se ajustaban demasiado bien a su teoría como para que pudieran

ser debidos al azar. Fisher usó el contraste de bondad de ajuste que había desarrollado Karl Pearson, para demostrar que la probabilidad de que los datos completos del experimento de Mendel se ajustaran a su teoría, al menos tan exactamente como Mendel había indicado, era igual a 0,00004.

Por ejemplo, Mendel publicó que, de 8023 guisantes de la segunda generación, 6022 resultaron ser amarillos y 2001 verdes. Esto es, el ratio experimental de guisantes verdes de segunda generación que obtuvo fue 2001/8023 = 0,2494, que casi coincide exactamente con 0,25, la probabilidad teórica. Aunque tan buen ajuste no es inverosímil en sí mismo (el que se produzca un ajuste experimental al menos tan bueno como el reportado por Mendel puede ocurrir en aproximadamente un 10% de los casos), el problema es que casi todos los experimentos indicados por Mendel mostraban ajustes inusualmente tan concordantes con las probabilidades teóricas como éste. La combinación de todos los experimentos registrados por Mendel llevó a Fisher a obtener la citada probabilidad (p valor) de 0,00004.

Aunque Fisher creyó que los datos de Mendel habían sido manipulados, no culpó de ello directamente a Mendel. De hecho, Fisher llegó a creer que posiblemente la manipulación se debió a que alguno de sus asistentes conocía los resultados que Mendel esperaba. (Naturalmente, también es plausible que el mismo Mendel cometiera errores al registrar los datos. Incluso la gente honesta puede ver lo que no está escrito cuando creen que debería estarlo.)

En la sección 13.2 se presenta el contraste de bondad de ajuste de la chi-cuadrado, que puede ser utilizado para determinar en qué medida un conjunto de datos dado se ajusta bien a un determinado modelo de probabilidad.

En la sección 13.3 se considerarán poblaciones en las que cada miembro está clasificado de acuerdo con dos características distintas. Se mostrará cómo se puede utilizar el contraste de bondad de ajuste para verificar la hipótesis de que las dos características de un miembro de la población elegido aleatoriamente son independientes.

Las dos características del citado miembro de la población serán independientes si el conocimiento de una de ellas no afecta a las probabilidades de la otra característica. Mientras que en la sección 13.3 se supone que los datos provienen de una muestra aleatoria extraída de la población total, en la sección 13.4 se considerará un tipo distinto de proceso de muestreo. Según éste, primero se centra la atención en una de las características. Después, para cada uno de los distintos valores posibles de dicha característica, se extrae una muestra aleatoria de la subpoblación cuyos miembros toman el valor citado. Por ejemplo, si una de las características es el sexo, en lugar de extraer una muestra aleatoria de la población al completo, tal como se hace en la sección 13.3, ahora se extraerán dos muestras aleatorias de las subpoblaciones de varones y hembras, respectivamente. Se presentará un contraste de independencia cuando se utilice este tipo de esquema de muestreo. Adicionalmente, se explicará cómo se pueden utilizar los resultados de la sección 13.4 para contrastar la hipótesis de que un número arbitrario de proporciones poblacionales son iguales. En el caso particular de que se trate de dos poblaciones, este contraste es idéntico al presentado en la sección 10.6.

13.2 Contraste de bondad de ajuste de la chi-cuadrado

Consideremos una gran población y supongamos que cada miembro de ella toma un valor que puede ser 1 o 2 o 3 o … o k. Para un conjunto dado de probabilidades p_i, $i = 1, \ldots, k$, se considerará el problema de cómo contrastar la hipótesis nula de que los p_i representen,

para cada i, la proporción de individuos de la población que toman el valor i. Es decir, si P_i denota las proporciones verdaderas en la población de los individuos que toman el valor i, para $i = 1, \ldots, k$, se pretenderá contrastar la hipótesis nula

$$H_0: P_1 = p_1, P_2 = p_2, \ldots, P_k = p_k$$

frente a la hipótesis alternativa

$$H_1: P_i \neq p_i \qquad \text{para algún } i,\ i = 1, \ldots, k$$

Ejemplo 13.1 Se sabe que un 41% de la población de Estados Unidos es del tipo sanguíneo A, que un 9% es del tipo B, que un 4% del tipo AB y que el 46% restante del tipo O. Supongamos que se sospecha que la distribución de los tipos de sangre entre la población de individuos que padece cáncer de estómago difiere de la distribución que afecta a la población total.

Para comprobar si la distribución de tipos sanguíneos es diferente entre aquellos que sufren cáncer de estómago se podría contrastar la hipótesis nula

$$H_0: P_1 = 0{,}41, P_2 = 0{,}09, P_3 = 0{,}04, P_4 = 0{,}46$$

donde, entre todos aquellos individuos que padecen cáncer de estómago, P_1 representa a la proporción de individuos con sangre del tipo A, P_2 es la proporción de individuos con sangre tipo B, P_3 es la proporción de aquellos que tienen la sangre tipo AB y P_4 representa a la proporción de individuos con el tipo sanguíneo O. Rechazar la hipótesis nula nos permitiría concluir que la distribución de tipos sanguíneos de los individuos que padecen cáncer de estómago difiere realmente de la distribución del total de la población.

En el escenario anterior, cada individuo de la población de pacientes de cáncer de estómago puede tomar uno de cuatro valores posibles, dependiendo de su grupo sanguíneo. Nuestra intención será contrastar la hipótesis de que $P_1 = 0{,}41$, $P_2 = 0{,}09$, $P_3 = 0{,}04$ y $P_4 = 0{,}46$ representan las proporciones de individuos que toman cada uno de los cuatro valores posibles dentro de esa población. ■

Para contrastar la hipótesis de que $P_i = p_i$, $i = 1, \ldots, k$ se precisa, en primer lugar, extraer una muestra aleatoria de individuos de la población. Supongamos que esta muestra es de tamaño n. Denotemos por N_i el número de individuos de la muestra que toman el valor i, para $i = 1, \ldots, k$. Si la hipótesis nula es cierta, cada individuo de la muestra tomará el valor i con una probabilidad p_i. Adicionalmente, puesto que se asume que la población es muy grande, se tiene que los sucesivos valores de los individuos de la muestra son independientes. Así pues, si la hipótesis nula es cierta, la distribución de N_i coincidirá con la del número de éxitos en n pruebas independientes, y p_i será la probabilidad de éxito en cada prueba. Es decir, si H_0 es cierta, N_i será una variable aleatoria binomial con parámetros n y p_i. Puesto que el valor esperado de una binomial coincide con el producto de sus parámetros, se tiene que, cuando H_0 es cierta,

$$E[N_i] = np_i \qquad i = 1, \ldots, k$$

Para cada i, denotemos por e_i el número esperado de resultados iguales a i, si se asume que H_0 es cierta. Esto es,

$$e_i = np_i$$

Así pues, si H_0 es cierta, se puede esperar que N_i esté relativamente próximo a e_i. Es decir, si la hipótesis nula es cierta, la magnitud $(N_i - e_i)^2$ no debería ser excesivamente grande en relación con e_i. Dado que esto es cierto para cada i, una forma razonable de de contrastar H_0 es la de computar el valor del estadístico del contraste

$$\text{TS} = \sum_{i=1}^{k} \frac{(N_i - e_i)^2}{e_i}$$

y rechazar H_0 cuando TS sea suficientemente grande.

Para determinar lo grande que debe ser TS para que esté justificado rechazar de la hipótesis nula se utilizará un resultado que fue demostrado por Karl Pearson en 1900. Este resultado establece que, para valores grandes de n, TS sigue aproximadamente una distribución chi-cuadrado con $k - 1$ grados de libertad. Denotemos por $\chi^2_{k-1,\,\alpha}$ el percentil de orden $100(1 - a)\%$ de esta distribución; esto significa que la probabilidad de que una variable aleatoria chi-cuadrado con $k - 1$ grados de libertad sobrepase este valor es igual a α (figura 13.3). En consecuencia, el contraste aproximado a nivel de significación a de la hipótesis nula H_0 frente a la hipótesis alternativa H_1 actuará como sigue:

Rechazar H_0	si $\text{TS} \geq \chi^2_{k-1,\,\alpha}$
No rechazar H_0	en otro caso

El contraste anterior se denomina *contraste de bondad de ajuste de la chi-cuadrado*. Para valores razonablemente grandes de n, el contraste anterior tiene un nivel de significación aproximadamente igual a α. Grosso modo, una regla empírica generalmente aceptada indica que, si n es suficientemente grande, la citada aproximación es buena si $e_i \geq 1$ para todo i y que al menos un 80% de los valores e_i son mayores que 5.

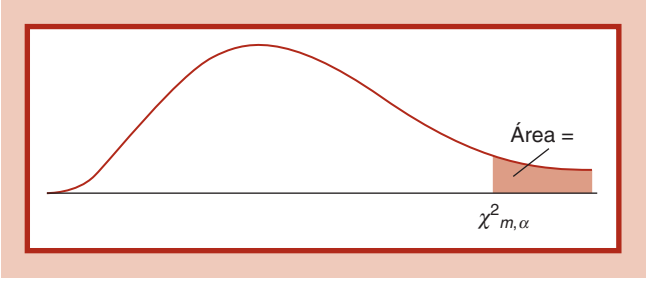


Figura 13.3 Percentil de la chi-cuadrado $P(\chi^2_m \geq \chi^2_{m,\alpha}) = \alpha$.

Tabla 13.1 Algunos valores de $\chi^2_{m,\alpha}$

m	$\alpha = 0{,}99$	$\alpha = 0{,}95$	$\alpha = 0{,}05$	$\alpha = 0{,}01$
1	0,000157	0,00393	3,841	6,635
2	0,0201	0,103	5,991	9,210
3	0,115	0,352	7,815	11,345
4	0,297	0,711	9,488	13,277
5	0,554	1,145	11,070	15,086
6	0,872	1,635	12,592	16,812
7	1,239	2,167	14,067	18,475

En la tabla D.3 del Apéndice se muestran los valores de $\chi_{m,\alpha}$ para distintos valores de m y α. Una parte de esta tabla está incluida en la tabla 13.1.

Ejemplo 13.2 Supongamos que, en el ejemplo 13.1, se ha extraído una muestra aleatoria de pacientes con cáncer de estómago en la que 92 pacientes resultaron ser del grupo sanguíneo A, 20 lo fueron del grupo B, 4 tienen sangre del tipo AB y 84 del grupo O. Al nivel de significación del 5%, ¿son estos datos lo suficientemente significativos para permitirnos rechazar la hipótesis nula de que la distribución de tipos sanguíneos entre los pacientes de cáncer de estómago es igual a la distribución de tipos sobre el total de la población?

Solución Las frecuencias observadas son

$$N_1 = 92 \quad N_2 = 20 \quad N_3 = 4 \quad N_4 = 84$$

mientras que las frecuencias esperadas, si se asume que H_0 es cierta, son

$$e_1 = np_1 = 200 \times 0{,}41 = 82$$
$$e_2 = np_2 = 200 \times 0{,}09 = 18$$
$$e_3 = np_3 = 200 \times 0{,}04 = 8$$
$$e_4 = np_4 = 200 \times 0{,}46 = 92$$

Así pues, el valor del estadístico del contraste es

$$\text{TS} = \frac{(92 - 82)^2}{82} + \frac{(20 - 18)^2}{18} + \frac{(4 - 8)^2}{8} + \frac{(84 - 92)^2}{92}$$
$$= 4{,}1374$$

Puesto que este valor no sobrepasa $\chi^2_{3,\,0{,}05} = 7{,}815$ (obtenido de la tabla 13.1) se sigue que, al nivel de significación del 5%, no se puede rechazar la hipótesis nula de que la distribu-

ción de tipos sanguíneos entre los pacientes de cáncer de estómago coincide con la distribución de tipos en toda la población. ■

El contraste de bondad de ajuste de la chi-cuadrado también se puede llevar a cabo determinando el *p* valor de los datos resultantes. Si, a partir de los datos, se obtiene un valor del estadístico del contraste igual a v, el *p* valor es igual a la probabilidad de que ocurra un valor al menos tan grande como v, asumiendo que H_0 es cierta. Ahora bien, cuando H_0 es cierta, la distribución del estadístico del contraste TS es aproximadamente una chi-cuadrado con $k - 1$ grados de libertad. En consecuencia, el *p* valor coincide aproximadamente con la probabilidad de que la variable aleatoria chi-cuadrado con $k - 1$ grados de libertad sea mayor o igual que v. Así pues, la hipótesis nula se debe rechazar a cualquier nivel de significación mayor o igual que el *p* valor, y no se debe rechazar a cualquier nivel de significación inferior a dicho *p* valor.

Para determinar el *p* valor del contraste de la chi-cuadrado

1. Calcule el valor del estadístico del contraste TS.
2. Si el valor de TS es v, el *p* valor viene dado por

$$p \text{ valor} = P\{\chi^2_{k-1} \geq v\}$$

donde χ^2_{k-1} es una variable aleatoria chi-cuadrado con $k - 1$ grados de libertad.

Se puede utilizar el Programa 13-1 para determinar tanto el valor del estadístico del contraste TS como el *p* valor resultante.

Ejemplo 13.3 Para determinar si los accidentes de trabajo ocurren con mayor probabilidad en ciertos días de la semana, se han recogido datos de todos los accidentes que han ocurrido en una fábrica de automóviles del norte de California y que han requerido atención médica. Los datos obtenidos afectan a 250 accidentes, distribuidos entre los días de la semana como sigue:

Lunes	62
Martes	47
Miércoles	44
Jueves	45
Viernes	52

Utilice los datos anteriores para contrastar, al nivel de significación del 5%, la hipótesis de que los accidentes laborales se dan con la misma probabilidad a lo largo de todos los días de la semana.

Solución Se pretende contrastar la hipótesis de que

$$P_i = \frac{1}{5} \qquad i = 1, 2, 3, 4, 5$$

Los datos observados son $N_1 = 62$, $N_2 = 47$, $N_3 = 44$, $N_4 = 45$ y $N_5 = 52$. Con el Programa 13-1 se obtiene que los valores resultantes del estadístico del contraste TS y del p valor son

$$\text{TS} = 4{,}36 \qquad p \text{ valor} = 0{,}359$$

Así pues, un valor de TS al menos tan grande como el obtenido se puede esperar que ocurra un 35,9% de las veces cuando H_0 es cierta; por consiguiente, no se puede rechazar la hipótesis nula de que los accidentes suceden con la misma probabilidad a lo largo de todos los días de la semana. ■

En ocasiones se hacen públicos conjuntos de datos que están tan de acuerdo con los valores esperados, bajo la hipótesis nula, que uno puede sospechar la posibilidad de que los datos hayan sido manipulados. Una forma de comprobar la verosimilitud de esta posibilidad consiste en calcular el valor del estadístico del contraste TS para determinar, después, en qué medida es probable haber obtenido un valor mayor o igual que v si se asume que la hipótesis nula es cierta. Esto es, uno debería determinar $P\{\chi^2_{k-1} \geq v\}$. Un valor extremadamente pequeño de esta probabilidad será una fuerte evidencia que confirme la posible manipulación de los datos.

Ejemplo 13.4 En la introducción de este capítulo se ha comentado un experimento efectuado por Gregor Mendel en el que se explicaba que, sobre 8023 cruces entre guisantes híbridos se habían obtenido 6022 guisantes amarillos y 2001 guisantes verdes. En teoría, de cada cruce se debería obtener un guisante amarillo con una probabilidad de 3/4, y uno verde con probabilidad 1/4. Para determinar si los datos se ajustan demasiado bien al modelo se empezará determinado el valor del estadístico del contraste TS.

Los parámetros de este problema son

$$n = 8023 \qquad k = 2 \qquad p_1 = 3/4 \qquad p_2 = 1/4 \qquad N_1 = 6022 \qquad N_2 = 2001$$

Puesto que

$$e_1 = 8023 \times \frac{3}{4} = 6017{,}25$$

$$e_2 = 8023 \times \frac{1}{4} = 2005{,}75$$

el valor del estadístico del contraste TS es

$$\text{TS} = \frac{(6022 - 6017{,}25)^2}{6017{,}25} + \frac{(2001 - 2005{,}75)^2}{2005{,}75}$$

$$= 0{,}015$$

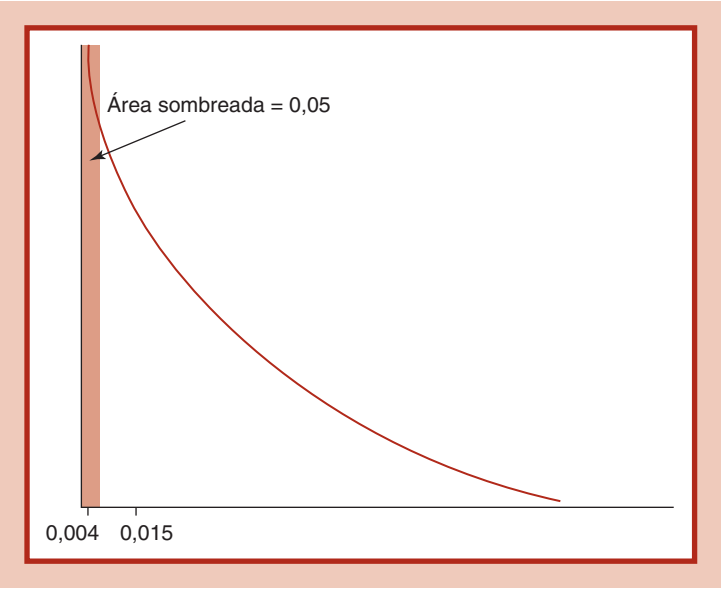


Figura 13.4 $\chi^2_{1,\,0,95} = 0{,}004$ implica que $P\{\chi^2_1 \leq 0{,}015\} \geq 0{,}05$.

Puesto que 0,015 es mayor que $\chi^2_{1,\,0,95} = 0{,}004$ se tiene que un valor menor o igual que 0,015 podría ocurrir más de un 5% de las veces (véase la figura 13.4). Por consiguiente, los datos no indican la existencia de manipulación.

De hecho, se puede ver (por ejemplo, mediante el Programa 13-2, que proporciona el *p* valor $P\{\chi^2_1 \geq 0{,}015\}$ como salida) que

$$P\{\chi^2_1 \leq 0{,}015\} = 0{,}0974$$

y, por consiguiente, aproximadamente un 10% de las veces el TS podría ser al menos tan pequeño como el valor obtenido con los datos de Mendel. Aunque esto por sí mismo no afirma la existencia de manipulación (consciente o inconsciente) de los datos, resulta que casi todos los conjuntos de datos difundidos por Mendel se ajustan tan bien como éste a las esperanzas teóricas. De hecho, la probabilidad de que la suma de los valores de todos los estadísticos del contraste de la chi-cuadrado publicados por Mendel, uno por cada experimento, fuera menor o igual que el valor obtenido usando los datos de Mendel es de 0,00004. ■

La tabla 13.2 sintetiza el contraste de bondad de ajuste de la chi-cuadrado.

Tabla 13.2 Contraste de bondad de ajuste de la chi-cuadrado.

Supongamos que cada miembro de una población puede tomar los valores 1, 2, . . . , k. Sea P_i la proporción de individuos de la población que toma el valor i, $i = 1, 2, \ldots, k$. Sea p_i, $i = 1, 2, \ldots, k$, un conjunto dado de valores no negativos cuya suma sea 1, y consideremos el contraste

$$H_0: P_i = p_i \qquad \text{para todo } i = 1, \ldots, k$$

frente a

$$H_1: P_i \neq p_i \quad \text{para algún } i = 1, \ldots, k$$

Sea $e_i = np_i$, $i = 1, 2, \ldots, k$, y hagamos n lo suficientemente grande como para que todos los e_i sean como mínimo 1 y para que al menos un 80% de ellos sean como mínimo 5.

Sea N_i el número de elementos de la muestra que toman el valor i. Utilice el estadístico del contraste

$$TS = \sum_{i=1}^{k} \frac{(N_i - e_i)^2}{e_i}$$

El contraste a nivel de significación α actuará como sigue

Rechazar H_0	si $TS \geq \chi^2_{k-1,\,\alpha}\chi^2_{k-1,\,\alpha}$
No rechazar H_0	en otro caso

Equivalentemente, si el valor de TS es v, el p valor viene dado por

$$p \text{ valor} = P\{\chi^2_{k-1} \geq v\}$$

Arriba, χ^2_{k-1} representa una variable aleatoria chi-cuadrado con $k - 1$ grados de libertad y $\chi^2_{k-1,\,\alpha}$ es el percentil de orden $100(1 - \alpha)\%$ de esta distribución.

Problemas

1. Determine los siguientes percentiles de la chi-cuadrado.

   (a) $\chi^2_{5,\,0,01}$ (b) $\chi^2_{5,\,0,05}$ (c) $\chi^2_{10,\,0,01}$

   (d) $\chi^2_{10,\,0,05}$ (e) $\chi^2_{20,\,0,05}$

2. Consideremos un conjunto de datos de 200 elementos con la siguiente tabla de frecuencias:

Resultado	Frecuencia
1	44
2	38
3	57
4	61


(Science Photo Library/ Photo Researchers)

Karl Pearson

Perspectiva histórica

Un suceso relevante en la historia de la Estadística ocurrió en 1900, cuando Karl Pearson publicó un artículo en la revista *The London, Edinburgh, and Dublín Philosophical Magazine and Journal of Science.* En ese artículo se presentaba por primera vez el contraste de bondad de ajuste de la chi-cuadrado. Resultó ser un suceso de gran importancia porque cambió la idea que hasta entonces se tenía de la Estadística. Hasta ese momento la mayor parte de los científicos veían la Estadística como una disciplina relacionada con la organización y la presentación de los datos, pero este artículo propició que muchos de ellos la visualizaran como una disciplina muy relacionada con los contrastes de hipótesis.

Considere el contraste de la hipótesis de que cada uno de los 200 valores de datos tiene la misma probabilidad de ser igual a cada uno de los enteros comprendidos entre 1 y 4.

(a) Indique las hipótesis nula y alternativa.

(b) Calcule el valor del estadístico del contraste.

(c) Al nivel de significación del 10% ¿qué conclusión se puede extraer?

(d) Repita el apartado (c) al nivel de significación del 5%.

(e) Repita el apartado (c) al nivel de significación del 1%.

3. En una determinada zona geográfica ha venido siendo aceptado históricamente que un 52% de los pacientes que van a los servicios de urgencia de los hospitales se encuentran en una condición estable, que un 32% se encuentran en una condición seria y que un 16% se encuentran en una condición crítica. Pese a ello, el gerente de un determinado hospital de la zona intuye que los porcentajes en su caso son diferentes. Para verificar su intuición, el gerente selecciona a una muestra aleatoria de 300 pacientes que han demandado los servicios de urgencia de su hospital en los pasados 6 meses. El número de pacientes muestrales que caen en cada grupo se muestra a continuación:

   Estables 148

   Serios 92

   Críticos 60

   ¿Prueban estos datos la sospecha del gerente? Explique detalladamente cuál es la hipótesis nula, y utilice un nivel de significación del 5%.

4. Los siguientes datos muestran cómo se distribuyen entre los días de la semana una muestra de 100 faltas de asistencia a clase de los estudiantes de un curso:

Día	Lunes	Martes	Miércoles	Jueves	Viernes
Fecuencia	27	19	13	15	26

   Contraste la hipótesis de que las faltas de asistencia a clase tienen la misma probabilidad a lo largo de todos los días de la semana. ¿Qué conclusiones se pueden extraer?

5. Considere un experimento con seis resultados posibles cuyas probabilidades se suponen que son 0,1, 0,1, 0,05, 0,4, 0,2 y 0,15. Se desea contrastar si estas probabilidades son las verdaderas mediante la realización de 60 repeticiones independientes del experimento. Si las veces en que se obtuvieron cada uno de los seis resultados posibles, sobre el total de repeticiones, fueron 4, 3, 7, 17, 16, y 13, ¿se debería rechazar la hipótesis? Utilice un nivel de significación del 5%.

6. En una determinada región, un 84% de los conductores no tiene accidentes a lo largo del año, un 14% tiene un solo accidente y un 2% tienen 2 accidentes o más. De una muestra aleatoria de 400 abogados, 308 no sufrieron ningún accidente, 66 tuvieron 1 accidente y 26 tuvieron 2 o más. Con estos datos, ¿se puede llegar a la conclusión de que los abogados no presentan el mismo perfil de accidentes que el resto de los conductores de la región?

7. La producción histórica de una máquina indica que cada unidad que produce es de

Calidad superior	con probabilidad 0,38
Calidad alta	con probabilidad 0,32
Calidad media	con probabilidad 0,26
Calidad baja	con probabilidad 0,04

   Una nueva máquina, diseñada para llevar a cabo el mismo trabajo, ha producido 500 unidades, con los resultados siguientes:

Calidad superior	222
Calidad alta	171
Calidad media	98
Calidad baja	9

   ¿Podría ocurrir que la diferencia en los resultados obtenidos con ambas máquinas se deba simplemente al azar? ¡Explique por qué!

8. Lance un dado 100 veces y registre las frecuencias obtenidas de cada uno de los seis resultados posibles. Utilice los datos resultantes para contrastar la hipótesis de que las seis caras del dado tienen la misma probabilidad de ocurrencia.

9. Un director de marketing mantiene que las ventas por correo tienen la misma probabilidad de que provengan de una cualquiera de cuatro regiones distintas. Un empleado no está de acuerdo y, por ello, ha recogido una muestra de 400 peticiones de venta. Con éstas se obtuvieron las siguientes cifras de ventas provenientes de cada región:

Región 1	106
Región 2	138
Región 3	76
Región 4	80

   Al nivel de significación del 5% ¿desaprueban estos datos la hipótesis del director? ¿Y si el nivel de significación fuera del 1%?

10. Se ha realizado un estudio para ver si los terremotos, al menos los de intensidad moderada (con valores mayores o iguales a 4,4 en la escala Richter), que tienen lugar al sur de California ocurren con mayor probabilidad en unos días de la semana que en otros. Los datos registrados sobre 1100 terremotos indican lo siguiente:

Día	Domingo	Lunes	Martes	Miércoles	Jueves	Viernes	Sábado
Número de terremotos	156	144	170	158	172	148	152

Al nivel de significación del 5%, contraste la hipótesis de que los terremotos tienen la misma probabilidad de que ocurran en cualquier día de la semana.

11. En las loterías de cierto Estado, los jugadores compran un boleto y sobre él seleccionan cuatro enteros comprendidos entre 0 y 36, inclusive. La comisión de loterías selecciona después cuatro números dentro del rango anterior de modo que todas las selecciones posibles son igualmente probables. Tras retirar el porcentaje de recaudación no repartida, que en ocasiones supone un 40%, la comisión de lotería divide la recaudación restante a partes iguales entre todos los jugadores que han acertado los cuatro números. Dado que todos los posibles grupos de los cuatro números tienen la misma probabilidad de ser seleccionados por la comisión, es fácil ver que es mejor que un jugador elija cuatro números "impopulares" de manera que, si gana, no tenga que compartir el premio con otros muchos ganadores. Esto es, puesto que la probabilidad de que un jugador gane, al igual que la cantidad a repartir, es igual cualquiera que sea la elección del jugador, es mejor que éste elija números que no tengan mucha probabilidad de ser elegidos por los otros jugadores.

    De este razonamiento surge la cuestión de si existen números que son realmente impopulares; esto es, números que se seleccionan menos frecuentemente que otros. Para responder a esta cuestión se podría llevar a cabo un contraste de la chi-cuadrado para comprobar si todas las elecciones posibles de la lotería son igualmente probables.

    Considere una lotería simplificada en la que cada jugador selecciona un entero comprendido entre 1 y 10. Supongamos que se ha extraído una muestra de aleatoria de 10 000 boletos de lotería comprados con anterioridad, de la que se han obtenido las frecuencias siguientes:

Número	**Frecuencia**
1	1122
2	1025
3	1247
4	818
5	1043
6	827
7	1149
8	946
9	801
10	1022

¿Estos datos prueban que los 10 números no se juegan con la misma frecuencia?

12. Los datos suministrados por la Oficina de Estadísticas de Trabajo de Estados Unidos indican que la población total de trabajadores de Estados Unidos en horario flexible se distribuye porcentualmente por clases se edad como sigue:

Clases de edad	Porcentaje
16–24	13,7
25–34	32,5
35–44	26,3
45–54	17,1
55 o más	10,4

Supongamos que en la ciudad de Sacramento se elige una muestra aleatoria de 240 trabajadores con horario flexible, en la que 24 tienen edades comprendidas entre 16 y 24 años, 94 están entre 25 y 34 años, 48 entre 35 y 44 años, 35 entre 45 y 54 años y 39 tienen 55 años o más. Al nivel de significación del 5%, ¿se puede concluir que la distribución por clases de edad de los trabajadores de Sacramento con horario flexible difiere de la distribución de la población total de dicho tipo de trabajadores?

13. El informe anual del año 1995 de las chicas scouts de Estados Unidos indica que un 59,8% de sus miembros tenían 8 o menos años de edad, que un 32,4% estaban entre los 9 y los 11 años y que un 7,8% tenían 12 o más años. En 2002 se extrajo una muestra aleatoria de 400 chicas scouts en la que 255 tenían 8 o menos años, 112 tenían una edad comprendida entre 9 y 11 años y que 33 tenían 12 o más años. Contraste la hipótesis de que los porcentajes de 2002 coinciden con los de 1995. Utilice un nivel de significación del 5%.

14. Karl Pearson indicó que él había lanzado 24 000 veces una moneda, resultando 12 012 veces cara y 11 988 veces cruz. ¿Es esto creíble? Explique el razonamiento en el que se basa la respuesta.

15. Los datos siguientes muestran los porcentajes de mujeres solteras por clases de edad que dieron a luz en 1986:

Clases de edad	Porcentaje
14 o menos	1,1
15–19	32,0
20–24	36,0
25–29	18,9
30 o más	12,0

Fuente: Centro Nacional de Estadísticas de la Salud de Estados Unidos, *Estadísticas de Vida de Estados Unidos.*

Una muestra reciente de 1000 nacimientos de mujeres no casadas indican que 42 de las madres tenían 14 años o menos, 403 tenían una edad comprendida entre los 15 y los 19 años, 315 tenían entre 20 y 24 años, 150 tenían entre 25 y 29 años, y 90 tenían 30 o más años. ¿Prueban estos datos que los porcentajes actuales difieren de los de 1986?

13.3 Contraste de independencia en poblaciones clasificadas de acuerdo con dos características

Consideremos una población grande en la que cada miembro se clasifica de acuerdo con dos características, designadas como característica X y característica Y. Supongamos que los posibles valores de la característica X se denotan por $1, 2, \ldots, r$ y que, análogamente, los valores posibles de la característica Y se denotan por $1, 2, \ldots, s$. Así pues, existen r posibles valores de la característica X y s posibles valores de la Y.

Ejemplo 13.5 Consideremos una población de adultos en edad de votar, y supongamos que cada adulto está clasificado de acuerdo con su sexo –mujer o varón– y su afinidad política –demócrata, republicano o independiente–. Sea X la característica que indica el sexo e Y la característica que representa la afinidad política. Puesto que existen dos posibles sexos y tres posibles afinidades políticas, $r = 2$ y $s = 3$. Hagamos que la característica X de una persona tome el valor 1 si se trata de una mujer y el valor 2 si se trata de un hombre. De igual forma, hagamos que la característica Y de una persona sea igual a 1 si es demócrata, 2 si es republicano y 3 si es independiente. Así pues, por ejemplo, una mujer que sea republicana tomará el valor 1 para la característica X y el valor 2 para la característica Y. ■

Denotemos por P_{ij} la proporción de elementos de la población cuya característica X toma el valor i y cuya característica Y toma el valor j, simultáneamente, siendo i cualquiera de los valores $1, 2, \ldots, r$ y j cualquiera de los valores $1, 2, \ldots, s$. Denotemos, además, como P_i a la proporción de elementos de la población cuyas características X toman el valor i, y como Q_j a la proporción de elementos de la población para los que su característica Y toma el valor j. Así pues, si X e Y denotan los valores de la característica X y de la característica Y de un miembro de la población elegido aleatoriamente, se tendrá que

$$P\{X = i, Y = j\} = P_{ij}$$
$$P\{X = i\} = P_i$$
$$P\{Y = j\} = Q_j$$

Ejemplo 13.6 Para la situación descrita en el ejemplo 13.5, P_{11} representará la proporción de mujeres de la población que de declaran demócratas, P_{12} será la proporción de mujeres republicanas y P_{13} representará la proporción de mujeres de la población que se declaran independientes. Las proporciones P_{21}, P_{22} y P_{23} son similares, sin más que cambiar mujeres por hombres. Los valores P_1 y P_2 representarán las proporciones poblacionales de mujeres y hombres, respectivamente; mientras que, por su parte, Q_1, Q_2 y Q_3 serán las proporciones poblacionales de aquellas personas con una tendencia demócrata, republicana e independiente, respectivamente. ■

Se pretende desarrollar un contraste de hipótesis que nos permita determinar si las características X e Y de un miembro de la población elegido aleatoriamente son independientes. Recordemos que X e Y son independientes si

$$P\{X = i, Y = j\} = P\{X = i\}P\{Y = j\}$$

de donde se sigue que nuestra intención será contrastar la hipótesis nula

$$\mathrm{H}_0\text{: } P_{ij} = P_i Q_j \quad \text{para todo } i = 1, \ldots, r,\ j = 1, \ldots, s$$

frente a la hipótesis alternativa

$$\mathrm{H}_1\text{: } P_{ij} \neq P_i Q_j \quad \text{para algunos valores de } i \text{ y } j$$

Para llevar a cabo este contraste de independencia se comenzará eligiendo una muestra aleatoria de elementos de la población de tamaño n. Denotemos por N_{ij} el número de elementos de la muestra que toman, simultáneamente, el valor i para la característica X y el valor j para la característica Y.

Ejemplo 13.7 Consideremos el ejemplo 13.5 y supongamos que se extrae una muestra aleatoria de 300 personas de la población, con los siguientes datos resultantes:

	j			
i	**Demócratas**	**Republicanos**	**Independientes**	**Total**
Mujeres	68	56	32	156
Hombres	52	72	20	144
Total	120	128	52	300

Por ejemplo, esto indica que la muestra aleatoria de tamaño 300 está formada por 68 mujeres con tendencia demócrata, por 56 mujeres con tendencia republicana y por 32 mujeres que se consideran independientes; esto es, $N_{11} = 68$, $N_{12} = 56$ y $N_{13} = 32$. De igual manera, $N_{21} = 52$, $N_{22} = 72$ y $N_{23} = 20$.

Esta tabla, que refleja el número de miembros de la muestra que caen dentro de cada una de las rs celdas posibles, se denomina *tabla de contingencia.* ■

Si la hipótesis de que las características X e Y de un miembro aleatoriamente elegido de la población es cierta, cada elemento de la muestra tendrá el valor i de la característica X y el valor j de la característica Y con una probabilidad $P_i Q_j$. De aquí se desprende, a partir de los resultados de la sección 13.2, que si estas probabilidades son conocidas se podría contrastar H_0 utilizando como estadístico del contraste

$$\mathrm{TS} = \sum_i \sum_j \frac{(N_{ij} - e_{ij})^2}{e_{ij}}$$

donde

$$e_{ij} = nP_iQ_j$$

La variable e_{ij} representa el número esperado, si se asume que H_0 es cierta, del número de elementos muestrales que toman simultáneamente los valores i y j para las características X e Y, respectivamente. Para el cómputo de TS se ha de calcular la suma de los términos de cada par i,j para cada uno de sus rs posibles valores. Si se asume que H_0 es cierta, TS seguirá aproximadamente una distribución chi-cuadrado con $rs - 1$ grados de libertad.

El problema de utilizar directamente este enfoque es que existen $r + s$ valores de P_i y Q_j, $i = 1, \ldots, r, j = 1, \ldots, s$, que no están especificados en la hipótesis nula. Así pues, necesitamos estimarlos previamente. Para hacerlo, denotemos por N_i y M_j el número de elementos muestrales que toman simultáneamente los valores i y j para las características X e Y, respectivamente. Puesto que N_i/n y M_j/n son las proporciones de los miembros de la muestra que, respectivamente, presentan el valor i para la característica X y el valor j para la característica Y, resulta natural utilizarlos como estimadores de P_i y Q_j. Esto es, se estimarán P_i y Q_j mediante

$$\hat{P}_i = \frac{N_i}{n} \qquad \hat{Q}_j = \frac{M_j}{n}$$

Esto conduce a utilizar el siguiente estimador de e_{ij}:

$$\hat{e}_{ij} = n\hat{P}_i\hat{Q}_j = \frac{N_iM_j}{n}$$

Dicho de otro modo, $\hat{e}_{ij}$ es igual al producto del número de miembros de la muestra que toman el valor i para la característica X (esto es, la suma de la fila i de la tabla de contingencia) por el número de miembros muestrales que toman el valor j para la característica Y (esto es, la suma de la columna j de la tabla de contingencia) dividido por el tamaño muestral n.

Así pues, parece razonable utilizar el siguiente estadístico para contrastar la independencia entre las características X e Y:

$$\text{TS} = \sum_i \sum_j \frac{(N_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}$$

donde los valores de $\hat{e}_{ij}$, $i = 1, \ldots, r, j = 1, \ldots, s$, se han especificado anteriormente.

Para determinar el conjunto de valores de TS para los que se debería rechazar la hipótesis nula, es necesario conocer la distribución de TS cuando la hipótesis nula es cierta. Se puede demostrar que la distribución del estadístico del contraste TS sigue aproximadamente una chi-cuadrado con $(r - 1)(s - 1)$ grados de libertad, si la hipótesis nula es cierta. De aquí se desprende que el contraste de H_0 a nivel de significación α actuará como sigue:

Rechazar H_0	si $\text{TS} \geq \chi^2_{(r-1)(s-1),\,\alpha}$
No rechazar H_0	en otro caso

Una observación técnica: No es difícil ver por qué el estadístico del contraste TS tiene $(r-1)(s-1)$ grados de libertad. Recordemos de la sección 13.2 que, si todos los valores P_i y Q_j se conocieran de antemano, el estadístico del contraste tendría $rs-1$ grados de libertad. (Esto es así porque k, el número de diferentes tipos de elementos de la población, es igual a rs.) Ahora bien, a primera vista puede parecer que se tengan que utilizar los datos para estimar $r+s$ parámetros. Sin embargo, dado que la suma de los P_i y de los Q_j ha de ser igual a 1 –esto es $\Sigma_i P_i = \Sigma_j Q_j = 1$– realmente se deben estimar $r-1$ de los P_i y $s-1$ de los Q_j. (Por ejemplo, si r es igual a 2, un estimador de P_1 proporciona automáticamente un estimador de P_2, ya que $P_2 = 1 - P_1$.) De aquí se desprende que realmente se precisan estimar $r-1+s-1 = r-s-2$ parámetros. Dado que se pierde un grado de libertad por cada parámetro estimado se desprende que el estadístico del contraste resultante tiene $rs-1-(r+s-2) = rs-r-s+1 = (r-1)(s-1)$ grados de libertad.

Ejemplo 13.8 Los datos del ejemplo 13.7 son los siguientes:

	j			
i	**1**	**2**	**3**	**Total** $= N_i$
1	68	56	32	156
2	52	72	20	144
Total $= M_j$	120	128	52	300

¿Qué conclusión se puede extraer? Utilice un nivel de significación del 5%.

Solución De los datos anteriores, los seis valores de

$$\hat{e}_{ij} = \frac{N_i M_j}{n}$$

son los siguientes:

$$\hat{e}_{11} = \frac{N_1 M_1}{n} = \frac{156 \times 120}{300} = 62{,}40$$

$$\hat{e}_{12} = \frac{N_1 M_2}{n} = \frac{156 \times 128}{300} = 66{,}56$$

$$\hat{e}_{13} = \frac{N_1 M_3}{n} = \frac{156 \times 52}{300} = 27{,}04$$

$$\hat{e}_{21} = \frac{N_2 M_1}{n} = \frac{144 \times 120}{300} = 57{,}60$$

$$\hat{e}_{22} = \frac{N_2 M_2}{n} = \frac{144 \times 128}{300} = 61{,}44$$

$$\hat{e}_{23} = \frac{N_2 M_3}{n} = \frac{144 \times 52}{300} = 24{,}96$$

El valor del estadístico del contraste es, pues,

$$\begin{aligned} \text{TS} &= \frac{(68 - 62{,}40)^2}{62{,}40} + \frac{(56 - 66{,}56)^2}{66{,}56} + \frac{(32 - 27{,}04)^2}{27{,}04} \\ &\quad + \frac{(52 - 57{,}60)^2}{57{,}60} + \frac{(72 - 61{,}44)^2}{61{,}44} + \frac{(20 - 24{,}96)^2}{24{,}96} \\ &= 6{,}433 \end{aligned}$$

Puesto que $r = 2$ y $s = 3$, $(r - 1)(s - 1) = 2$ y, por tanto, el valor de TS se debe comparar con el valor crítico $\chi^2_{2,\,0{,}05}$. De la tabla 13.1,

$$\chi^2_{2,\,0{,}05} = 5{,}991$$

Puesto que TS $\geq 5{,}991$, se debe rechazar la hipótesis nula al nivel de significación del 5%. Esto es, no se puede aceptar la hipótesis de que el sexo y la afinidad política sean independientes, al nivel de significación del 5%. ■

El contraste de la hipótesis de que las características X e Y de un miembro de la población elegido aleatoriamente son independientes puede también llevarse a cabo determinando el p valor de los datos. Esto se consigue calculando primero el valor del estadístico del contraste TS. Si este valor es n, el p valor viene dado por

$$p \text{ valor} = P\{\chi^2_{(r-1)(s-1)} \geq v\}$$

donde $\chi^2_{(r-1)(s-1)}$ representa una variable aleatoria chi-cuadrado con $(r - 1)(s - 1)$ grados de libertad.

El Programa 13-2 calcula el valor del estadístico del contraste y, después, determina el p valor resultante. El programa asume que los datos han sido presentados en forma de tabla de contingencia y pide al usuario que introduzca las sucesivas filas de esa tabla.

Ejemplo 13.9 Un científico de salud pública desea saber si existe una relación entre el estado civil de los pacientes tratados de depresión y la severidad de sus enfermedades. El científico seleccionó una muestra aleatoria de 159 pacientes que habían sido tratados de depresión en una clínica de salud mental, y los clasificó de acuerdo con el nivel de severidad de sus depresiones –severo, normal, y ligero– y con su estado civil. Resultaron los siguientes datos:

	Estado civil			
Estado depresivo	**Casado**	**Soltero**	**Viudo o divorciado**	**Total**
Severo	22	16	19	57
Normal	33	29	14	76
Ligero	14	9	3	26
Total	69	54	36	159

Determine el p valor del contraste de la hipótesis de que el nivel de depresión de los pacientes de la clínica es independiente de su estado civil.

Solución Se ha utilizado el Programa 13-2 para obtener el valor del estadístico del contraste y el p valor resultante, que coinciden con

$$\text{TS} = 6{,}828 \qquad p \text{ valor} = 0{,}145 \quad \blacksquare$$

El contraste de independencia se encuentra resumido en la tabla 13.3.

Tabla 13.3 Contrastaste de la independencia entre dos características de los miembros de una población.

Supongamos que cada miembro de una población presenta dos características X e Y. Denotemos por r y s el número de valores posibles que pueden tener las características X e Y, respectivamente. Para contrastar

H_0: las características de un miembro de la población seleccionado aleatoriamente son independientes

frente a

H_1: las características de un miembro seleccionado aleatoriamente de la población no son independientes

elijamos una muestra aleatoria de n miembros de la población. Denotemos por N_{ij} el número de elementos muestrales que simultáneamente presentan el valor i para la característica X y el valor j para la característica Y. Denotemos también por

$$N_i = \sum_j N_{ij} \qquad \text{y} \qquad M_j = \sum_i N_{ij}$$

el número de miembros de la muestra que presentan el valor i de la característica X y el valor j de la característica Y, respectivamente. El estadístico del contraste es

$$\text{TS} = \sum_i \sum_j \frac{(N_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}$$

donde $\hat{e}_{ij} = N_i M_j / n$. El contraste a nivel de significación α actuará como sigue:

Rechazar H_0	si $\text{TS} \geq \chi^2_{(r-1)(s-1),\alpha}$
No rechazar H_0	en otro caso

Equivalentemente, si el valor de TS es v, el p valor coincide con

$$p \text{ valor} = P\{\chi^2_{(r-1)(s-1)} \geq v$$

Problemas

1. La siguiente tabla de contingencia presenta los datos de una muestra extraída de una población clasificada mediante dos características:

	Característica *X*		
Característica *Y*	**A**	**B**	**C**
1	32	12	40
2	56	48	60

   (a) Determine el valor del estadístico del contraste cuando se intenta contrastar si las dos características son independientes.

   (b) Al nivel de significación del 5%, ¿se debería rechazar la hipótesis nula?

   (c) ¿Y con nivel de significación del 1%?

2. Existe cierta evidencia de que poseer un perro puede tener cierto valor de predicción cuando se intenta determinar si un individuo sobrevivirá a un ataque de corazón. Los datos siguientes proceden de una muestra de 95 individuos que sufrieron un ataque de corazón severo. Los datos clasifican a cada uno de estos individuos con respecto a (1) si continuaban o no con vida un año después del ataque sufrido, y (2) si tenían o no un perro en casa.

	Tenían perro	**No tenían perro**
Sobrevivieron	28	44
No sobrevivieron	8	15

(Photoworld/FPG)

Florence Nightingale

Perspectiva histórica

Su biógrafo definió a Florence Nightingale como la *estadística pasional*. Es una descripción apropiada de esta mujer, que casi definitivamente convirtió la enfermería en ciencia. Durante la guerra de Crimea, recogió y registró datos sobre las condiciones sanitarias y las tasas de mortalidad en los hospitales militares, y utilizó esos datos para demostrar que ambas muestras de datos eran estadísticamente dependientes. Su trabajo fue decisivo para mejorar las condiciones higiénicas de los hospitales, con lo que se pudieron salvar un gran número de vidas.

Florence Nightingale fue una seguidora del estadístico belga Adolphe Quetelet; ella creía, como él, que *"los accidentes ocurren con una regularidad asombrosa cuando existen unas condiciones similares"*. Sostenía que los administradores con éxito eran aquellos que tenían en cuenta los datos. También mantenía que el universo evolucionaba de acuerdo con un plan divino y que era trabajo de cada persona aprender a vivir en armonía con ello. Sin embargo, para entender dicho plan, creía que uno debía estudiar Estadística. En palabras de Karl Pearson: *"Para Florence Nightingale, la Estadística era más que una ciencia, era su religión."*

¿Prueban estos datos, al nivel de significación del 5%, que tener un perro y sobrevivir a un ataque de corazón son dependientes? Explique cuidadosamente cuál es la hipótesis nula y qué estadístico del contraste se utiliza.

3. Se selecciona una muestra de 187 votantes a quienes se les sugiere que evalúen la actuación del presidente de Estados Unidos en sus primeros 100 días de mandato. Utilice los datos resultantes para contrastar la hipótesis de que la evaluación de un individuo no depende de si éste es un hombre o una mujer.

	Mujeres	Hombres
Evaluación positiva	54	47
Evaluación negativa	20	32
No está seguro	23	11

Utilice un nivel de significación del 5%.

4. Una compañía de seguros pretende determinar si existe relación entre la frecuencia de accidentes de automóviles y el consumo de cigarrillos. Se extrajo una muestra aleatoria de 597 asegurados, con la que se obtuvieron los resultados siguientes:

Número de accidentes en los últimos 2 años	Fumadores	No fumadores
0	35	170
1	79	190
2 o más	57	66

Al nivel de significación del 5%, contraste la hipótesis de que la frecuencia de accidentes de un asegurado elegido aleatoriamente es independiente de si éste fuma o no fuma.

5. La dirección de un determinado hotel pretende saber si a todos los clientes se les trata con la misma consideración con independencia de los precios de sus habitaciones. Para ello se seleccionó una muestra aleatoria de 155 clientes recientes y se les preguntó sobre el servicio que habían recibido durante su estancia en el hotel. Resultaron los siguientes datos:

	Tipo de habitación		
Calificación del servicio	Económica	Estándar	De lujo
Excelente	30	21	9
Bueno	36	29	8
Aceptable	12	8	2

¿Qué conclusiones se pueden extraer?

6. Los siguientes datos clasifican una selección de profesores de una determinada universidad de acuerdo con su capacidad docente (evaluada por los estudiantes que asistieron a sus clases en el semestre anterior) y el número de cursos impartían en aquel semestre

	Número de cursos		
Capacidad docente	**1**	**2**	**3 o más**
Por encima de la media	12	10	4
Media	32	40	38
Por debajo de la media	7	12	25

Al nivel de significación del 5%, contraste la hipótesis de que la capacidad docente de un profesor es independiente del número de cursos que él o ella imparta.

7. En el problema 6, es posible que solamente ciertos profesores, por lo general aquellos que se especializan en la investigación, impartan un solo curso por semestre. Los cursos impartidos por este tipo de profesores tienden a ser más avanzados y a tener menos estudiantes que el resto de los cursos. Por este motivo, para saber si el número de cursos afecta a la capacidad docente, podría ser razonable considerar los datos del problema 6 tras haber borrado la columna relativa a aquallos profesores que imparten un solo curso. Haga este cambio y repita el problema 6.

8. El estado socioeconómico de los residentes de una determinada zona urbana se puede clasificar como de clase baja o bien como de clase media. A los miembros de una muestra de residentes se les preguntó sobre su actitud acerca de un plan de construcción de una clínica de salud pública en la zona. Los resultados fueron los siguientes:

	Clase socioeconómica	
Actitud	**Baja**	**Media**
A favor	87	63
En contra	46	55

Contraste la hipótesis, al nivel de significación del 5%, de que los residentes de las clases baja y media muestran una misma actitud hacia el plan de construcción de la clínica.

9. Una empresa de investigación de mercado ha enviado muestras de un nuevo champú a un cierto número de individuos. Los datos siguientes reflejan las opiniones de éstos sobre la calidad del champú, diferenciando a los encuestados por clases de edad.

	Grupo de edad (años)		
Calidad del champú	**15-20**	**21-30**	**Más de 30**
Excelente	18	20	41
Buena	25	27	43
Aceptable	17	15	26
Mala	3	2	8

¿Prueban estos datos que las opiniones sobre el champú difieren por clases de edad? Utilice un nivel de significación del 5%.

10. A partir de una muestra de pacientes de una determinada clínica sanitaria se obtuvieron los siguientes datos que se refieren a los hábitos de tabaco y a los niveles de colesterol en sangre de los pacientes encuestados:

	Niveles de colesterol en sangre		
Hábitos de tabaco	**Bajo**	**Moderado**	**Alto**
Fumador pertinaz	6	14	24
Fumador moderado	12	23	15
No fumador	23	32	11

(a) Al nivel de significación del 5%, ¿se debe rechazar la hipótesis de independencia entre los niveles de colesterol en sangre y los hábitos de tabaco?

(b) Repita el apartado (a), y en esta ocasión utilice un nivel de significación del 1%.

(c) ¿Implican los resultados obtenidos que una reducción en el consumo de tabaco produce una disminución en el nivel de colesterol en sangre? ¡Explique por qué!

11. Para ver si existe dependencia entre trabajo profesional de una persona y sus creencias religiosas se seleccionó una muestra aleatoria de 638 individuos de la población formada por el conjunto total de médicos, abogados e ingenieros en activo. Los resultados muestrales aparecen reflejados en la siguiente tabla de contingencia:

	Médicos	**Abogados**	**Ingenieros**
Protestantes	64	110	152
Católicos	60	86	78
Judíos	57	21	10

Contraste la hipótesis, al nivel de significación del 5%, de que la profesión y las creencias religiosas de los individuos citados son independientes. Repita lo anterior al nivel del 1%.

12. Observe la siguiente tabla de contingencia e intuya (sin efectuar cálculo alguno) cuál será el resultado de contrastar, al nivel de significación del 5%, la hipótesis de que las dos características de los datos son independientes.

	A	B	C
1	26	44	30
2	14	30	25
3	30	45	33

A continuación, realice los cálculos necesarios.

13. Repita el problema 11, tras multiplicar por dos todos los valores de datos.

14. Repita el problema 12, después de multiplicar igualmente por dos todos los valores de datos.

13.4 Contraste de independencia en tablas de contingencia con totales marginales fijos

En el ejemplo 13.5 se pretendía averiguar si, en una población determinada, el sexo y la tendencia política de sus miembros eran o no dependientes. Para contrastar esta hipótesis, primero se seleccionaba a una muestra aleatoria de individuos de la población total, y después se anotaban las características de cada individuo. Sin embargo, existe otra forma de recoger la información que consiste en fijar de antemano el número de hombres y mujeres que van observarse y, luego, se seleccionan dos muestras aleatorias de las subpoblaciones de hombres y mujeres, respectivamente, con unos tamaños iguales a los números fijados de antemano. Esto es, en lugar de que el número de hombres y mujeres se determinen al azar, uno podría decidir que tales números se fijen por adelantado. Dado que, cuando se muestrea de este modo, el número de mujeres y hombres observados coinciden con los valores previamente elegidos se dice que la tabla de contingencia resultante tiene *marginales fijos* (puesto que los totales se reflejan en los márgenes de la tabla).

Aunque los datos se recojan de esa forma, se puede utilizar el mismo contraste de hipótesis dado en la sección 13.3 para contrastar la independencia entre las dos características. El estadístico del contraste continúa siendo

$$\text{TS} = \sum_i \sum_j \frac{(N_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}$$

donde

N_{ij} = número de elementos de la muestra con valor i de la característica X y con valor j de la característica Y

N_i = número de elementos de la muestra con valor i de la característica X

M_j = número de elementos de la muestra con valor i de la característica Y

y

$$\hat{e}_{ij} = \frac{N_i M_j}{n}$$

siendo n el tamaño total de la muestra.

También se verifica de nuevo que, si H_0 es cierta, TS sigue aproximadamente una distribución chi-cuadrado con $(r - 1)(s - 1)$ grados de libertad. (Los valores r y s coinciden, como antes, con los números de los posibles valores de las características X e Y, respectivamente.) En otras palabras, el contraste de la hipótesis de independencia no se ve afectado por el hecho de que los totales marginales se hayan fijado de antemano o se hayan obtenido tras haber muestreado sobre la población al completo.

Ejemplo 13.10 Un grupo de 20 000 no fumadores y otro de 10 000 fumadores se sometieron a un seguimiento durante 10 años. Los datos siguientes reflejan el número de individuos de cada grupo que desarrollaron un cáncer de pulmón a lo largo del periodo de seguimiento.

	Fumadores	No fumadores	Total
Con cáncer de pulmón	62	14	76
Sin cáncer de pulmón	9 938	19 986	29 924
Total	10 000	20 000	30 000

Contraste la hipótesis de que fumar y desarrollar un cáncer de pulmón son independientes. Utilice un nivel de significación del 1%.

Solución Los elementos muestrales que caen dentro de cada celda ij son

$$\hat{e}_{11} = \frac{(76)(10\,000)}{30\,000} = 25{,}33$$

$$\hat{e}_{12} = \frac{(76)(20\,000)}{30\,000} = 50{,}67$$

$$\hat{e}_{21} = \frac{(29\,924)(10\,000)}{30\,000} = 9974{,}67$$

$$\hat{e}_{22} = \frac{(29\,924)(20\,000)}{30\,000} = 19\,949{,}33$$

Así pues, el valor del estadístico del contraste es

$$\begin{aligned} \text{TS} &= \frac{(62 - 25{,}33)^2}{25{,}33} + \frac{(14 - 50{,}67)^2}{50{,}67} + \frac{(9938 - 9974{,}67)^2}{9974{,}67} \\ &\quad + \frac{(19\,986 - 19\,949{,}33)^2}{19\,949{,}33} \\ &= 53{,}09 + 26{,}54 + 0{,}13 + 0{,}07 = 79{,}83 \end{aligned}$$

Puesto que este valor es mucho mayor que $\chi^2_{1,0,01} = 6{,}635$ se debe rechazar la hipótesis nula de que una persona elegida aleatoriamente desarrolla o no cáncer de pulmón con independencia de si es o no fumadora. ■

Se verá a continuación cómo se puede utilizar todo lo anteriormente visto en esta sección para contrastar la hipótesis de que m proporciones poblacionales son iguales. Para empezar, consideremos m poblaciones distintas de individuos. Supongamos que p_i es la proporción de miembros de la i-ésima población que están a favor de una determinada propuesta, y consideremos el contraste cuya hipótesis nula es que todos los p_i son iguales. Es decir, se querrá contrastar

$$H_0: p_1 = p_2 = \cdots = p_m$$

frente a

$$H_1: \text{no todos los } p_i \text{ son iguales}$$

Para llevar a cabo el contraste anterior consideremos primero la superpoblación compuesta por el conjunto total de miembros de las m poblaciones distintas. Cualquier individuo de esta superpoblación puede ser clasificado de acuerdo con dos características. La primera de ellas especificará a cuál de las m poblaciones pertenece el individuo en cuestión, y la segunda indicará si el individuo está o no a favor de la propuesta citada. El contraste coincide, pues, con plantear si para los miembros de la superpoblación la característica de los valores que indican si se está a favor o en contra de la propuesta es independiente de la característica que representa a qué población pertenece cada individuo. Es decir, la hipótesis nula H_0 equivale a la hipótesis de independencia entre las dos características de la superpoblación.

Por consiguiente, se puede contrastar H_0 si se seleccionan muestras independientes con tamaños fijados de antemano de cada una de las m poblaciones citadas. Si se denota como M_i el tamaño de la muestra extraída de la población i, para $i = 1, \ldots, m$, se puede contrastar H_0 mediante el contraste de la independencia en la siguiente tabla de contingencia:

	Población				
	1	**2**	**...**	***m***	**Total**
A favor	F_1	F_2	...	F_m	N_1
En contra	A_1	A_2	...	A_m	N_2
Total	M_1	M_2	...	M_m	

En esta tabla, F_i y A_i representan los números de individuos muestrales de la población i que están a favor y en contra de la propuesta, respectivamente.

Ejemplo 13.11 En un estudio reciente se indicaba que se habían encuestado a 500 secretarias seleccionadas aleatoriamente en los cuatro países a que hacía referencia el estudio.

Una de las preguntas de la encuesta estaba relacionada con si las secretarias sentían haber sufrido abuso verbal o sexual en su trabajo. Los datos resultantes fueron los siguientes:

País	Número de secretarias que sufrieron abuso
Australia	28
Alemania	30
Japón	58
Estados Unidos	55

Basándonos en estos datos, ¿es plausible mantener que las proporciones de secretarias que creen haber sufrido abuso en su puesto de trabajo son iguales en los cuatro países contemplados en el estudio?

Solución Poniendo los datos en forma de una tabla de contingencia, resulta lo siguiente:

	País				
	1	**2**	**3**	**4**	**Total**
Sometidas a abuso	28	30	58	55	171
No sometidas a abuso	472	470	442	445	1829
Total	500	500	500	500	2000

Se puede contrastar la hipótesis nula contrastando la independencia de las características de la anterior tabla de contingencia. Si se utiliza el Programa 13-2, el valor del estadístico del contraste y el p valor resultante son

$$\text{TS} = 19{,}51 \qquad p \text{ valor} = 0{,}0002$$

En consecuencia, la hipótesis de que los porcentajes de las secretarias que creen haber sufrido abuso en su trabajo son iguales en los cuatro países se debe rechazar al nivel de significación del 1% (y, de hecho, a cualquier nivel de significación superior al 0,02%). ■

Cuando existen solamente dos poblaciones, el contraste anterior de igualdad de proporciones poblacionales coincide exactamente con el presentado en la sección 10.6.

Problemas

1. De los resultados del ejemplo 13.10, ¿se puede concluir que fumar causa cáncer de pulmón? ¿Qué otras explicaciones son posibles?
2. En un estudio sobre las preferencias de escolarización y las rentas de las familias se entrevistaron a 100 familias de renta alta y a 100 de renta baja en una determinada ciu-

dad. Cada familia encuestada informó sobre qué tipo de escuela prefería para sus hijos. Los datos resultantes fueron los siguientes:

Preferencia	Renta alta	Renta baja
Pública	22	19
Privada seglar	31	39
Privada no seglar	47	42

¿Qué conclusiones se pueden extraer?

3. Han sido sometidas a seguimiento dos muestras durante un año, una de 300 automóviles que disponían de teléfonos móviles y otra de 400 coches sin tales teléfonos. La tabla siguiente muestra los coches de cada muestra que sufrieron accidentes a lo largo de dicho año.

	Con accidentes	Sin accidentes
Con teléfono móvil	22	278
Sin teléfono móvil	26	374

Utilice estos datos para contrastar la hipótesis de que tener un teléfono móvil en el coche y sufrir accidentes son independientes. Considere un nivel de significación del 5%.

4. Una cadena de prensa escrita muestreó a 100 lectores de sus tres periódicos más importantes para determinar la clase económica de cada uno. Los resultados obtenidos se muestran a continuación:

	Periódico		
Economic class	1	2	3
Media baja	22	25	28
Media	41	37	44
Media alta	37	38	28

Contraste la hipótesis de que el periódico que lee un individuo es independiente de la clase económica del mismo. Utilice un nivel de significación del 5%.

5. La tabla siguiente muestra el número de piezas defectuosas y no defectuosas presentes en dos muestras de piezas producidas antes y después de haber introducido una modificación en el proceso de producción.

	Defectuosas	No defectuosas
Antes	22	404
Después	18	422

¿Prueban estos datos que la modificación ha ocasionado un cambio en el porcentaje de piezas defectuosas producidas?

6. Han sido seleccionados aleatoriamente 100 estudiantes de un curso de Estadística con 200 alumnos para que siguieran las clases por televisión en lugar de asistir en persona, tal como hicieron los restantes 100 alumnos. Las calificaciones finales del total de alumnos fueron las siguientes:

	Sobresaliente	**Notable**	**Aprobado**	**Suspenso**
Alumnos presenciales	22	38	35	5
Alumnos por televisión	18	32	40	10

Contraste la hipótesis de que las calificaciones finales son independientes de si el alumno ha seguido las clases en persona o por televisión. Al nivel de significación del 5%, ¿se puede rechazar la hipótesis de independencia? ¿Y al nivel del 1%?

7. Para estudiar el efecto que tiene el agua fluorada sobre el deterioro de la dentadura, se seleccionaron dos comunidades con el mismo nivel socioeconómico, aproximadamente. En una de las comunidades se disponía de agua fluorada, mientras que en la otra no. De cada una de estas comunidades se extrajo una muestra aleatoria de 200 adolescentes y se observó el número de caries que tenía cada uno. Los resultados obtenidos fueron los siguientes:

Caries	**Con agua fluorada**	**Sin agua fluorada**
0	154	133
1	20	18
2	14	21
3 o más	12	28

¿Prueban estos datos que, al nivel de significación del 5%, el número de caries dentales no es independiente de si el agua suministrada es fluorada o no? ¿Qué ocurre al nivel del 1%?

8. Un comerciante de automóviles ha enviado tarjetas postales a 990 clientes potenciales, en las que les ofrece el que puedan probar gratis uno de sus coches. Cada tarjeta postal tenía un color que podía ser rojo, blanco, azul claro o verde. A continuación se muestra el número de clientes que respondieron y el color de las tarjetas que recibieron:

	Roja	**Blanca**	**Azul**	**Verde**
Respondieron	108	106	105	127
No respondieron	142	144	135	123

Contraste la hipótesis de que el color de la tarjeta enviada no afecta a la capacidad de responder del receptor. Utilice un nivel de significación del 5%.

9. Se seleccionaron muestras de 50 estudiantes universitarios, 40 profesores de universidad y 60 trabajadores de banca, y después se observó quiénes fumaban y quiénes no. A continuación se muestran los resultados obtenidos:

Grupo	Número de no fumadores
Estudiantes universitarios	18
Profesores de universidad	12
Trabajadores de banca	24

(a) Al nivel de significación del 10%, contraste la hipótesis de que los colectivos de estudiantes universitarios, de profesores de universidad y de trabajadores de banca presentan los mismos porcentajes de fumadores.

(b) Repita el apartado (a) al nivel de significación del 5%.

(c) Repítalo también al nivel de significación del 1%.

10. Para determinar si las demandas judiciales por errores médicos se interponen con mayor probabilidad tras determinados tipos de operaciones que tras otros se seleccionaron y analizaron muestras de tres tipos de operaciones, y resultaron los datos siguientes:

Tipo de operación	Número de operaciones muestreadas	Número de operaciones con demanda judicial
Cirugía de corazón	400	16
Cirugía cerebral	300	19
Apendicitis	300	7

Contraste la hipótesis de que los porcentajes de las operaciones de cirugía que fueron demandadas judicialmente son iguales para los tres tipos citados.

(a) Utilice un nivel de significación del 5%.

(b) Utilice un nivel de significación del 1%.

Términos clave

Contraste de bondad de ajuste: Contraste estadístico de la hipótesis de que un determinado conjunto de k probabilidades representa la proporción de elementos de una población grande que caen dentro de k categorías distintas.

Tabla de contingencia: Tabla que clasifica cada elemento de una muestra de acuerdo con dos características distintas.

Resumen

Contraste de bondad de ajuste: Consideremos una población grande de elementos, cada uno de los cuales puede tomar un valor igual a 1 o 2 o . . . o k. Denotemos por P_i la proporción de elementos de la población que toman el valor i, $i = 1, \ldots, k$. Para un conjunto dado de probabilidades $p_1, \ldots, p_k$ ($p_i \geq 0, \sum_i p_i = 1$), consideremos el contraste de la hipótesis

$$H_0: P_i = p_i \qquad \text{para todo } i = 1, \ldots, k$$

frente a la hipótesis alternativa

$$H_1: P_i \neq p_i \qquad \text{para algún } i, i = 1, \ldots, k$$

Para llevar a cabo el contraste se seleccionará primero una muestra aleatoria de n elementos de la población. Denotemos por N_i el número de elementos de la muestra que toman el valor i. El estadístico del contraste que debe emplearse es

$$\text{TS} = \sum_{i=1}^{k} \frac{(N_i - e_i)^2}{e_i}$$

donde

$$e_i = np_i$$

Si H_0 es cierta, e_i es igual al número esperado de elementos de la muestra que toman el valor i.

Para llevar a cabo el contraste se rechazará H_0 si el valor de TS es suficientemente grande. Para determinar lo grande que debe ser este valor se utiliza el hecho de que, cuando H_0 es cierta, TS se distribuye aproximadamente como una chi-cuadrado con $k - 1$ grados de libertad. Esto implica que el contraste a nivel de significación α actuará como sigue:

Rechazar H_0	si $\text{TS} \geq \chi^2_{k-1,\alpha}$
No rechazar H_0	en otro caso

El valor $\chi^2_{k-1,\alpha}$ se define como

$$P\{\chi^2_{k-1} \geq \chi^2_{k-1,\alpha}\} = \alpha$$

donde χ^2_{k-1} representa una variable aleatoria chi-cuadrado con $k - 1$ grados de libertad.

Este contraste se denomina *contraste de bondad de ajuste de la chi-cuadrado*. Puede igualmente llevarse a cabo calculando el p valor del conjunto de datos. Si el valor observado de TS es ν, el p valor es

$$p \text{ valor} = P\{\chi^2_{k-1} \geq \nu\}$$

Se puede usar el Programa 13-1 para obtener tanto el valor del estadístico del contraste como el p valor resultante.

Contraste de independencia en poblaciones cuyos elementos están clasificados de acuerdo con dos características distintas: Supongamos ahora que cada elemento de una población está clasificado de acuerdo con dos características distintas, que denominaremos característica X y característica Y. Supongamos que los valores posibles de la característica X son $1, 2, \ldots, r$, y que los valores posibles de la característica Y son $1, 2, \ldots, s$. Denotemos por P_{ij} la proporción de elementos de la población que tienen el valor i de la característica X y el valor j de la característica Y. Igualmente, denotemos por P_i la proporción de miembros de la población que toman el valor i para la característica X, $i = 1, \ldots, r$; y denotemos por Q_j la proporción de elementos de la población que toman el valor j para la característica Y, $j = 1, \ldots, s$.

Supongamos que se pretende contrastar como hipótesis nula que las características X e Y de un miembro elegido aleatoriamente de la población son independientes. Es decir, consideremos el contraste de la hipótesis nula

$$\mathrm{H}_0\colon P_{ij} = P_i Q_j \qquad \text{para todo } i, j$$

frente a la hipótesis alternativa

$$\mathrm{H}_1\colon P_{ij} \neq P_i Q_j \qquad \text{para algún } i, j$$

Para contrastar estas hipótesis, se extraerá una muestra de tamaño n de la población. Denotemos por N_{ij} el número de elementos muestrales que presentan el valor i de la característica X y el valor j de la característica Y, simultáneamente. Igualmente, denotemos por N_i el número de elementos de la muestra que presentan el valor i de la característica X, y denotemos por M_j el número de elementos de la muestra que presentan el valor j de la característica Y. El estadístico del contraste utilizado para contrastar H_0 es

$$\mathrm{TS} = \sum_i \sum_j \frac{(N_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}$$

donde

$$\hat{e}_{ij} = \frac{N_i M_j}{n}$$

El sumatorio que aparece en la expresión de TS se extiende a todos los posibles rs valores del par ij. El contraste a nivel de significación a actuará como sigue:

Rechazar H_0	si $\mathrm{TS} \geq \chi^2_{(r-1)(s-1),\,\alpha}$
No rechazar H_0	en otro caso

De modo equivalente se puede llevar a cabo el contraste calculando el p valor. Si el valor de TS es v, el p valor viene dado por

$$p \text{ valor} = P\{\chi^2_{(r-1)(s-1)} \geq v\}$$

Se puede usar el Programa 13-2 para obtener tanto el valor de TS como el p valor resultante.

Exactamente el mismo contraste anterior se puede utilizar cuando no se elige una muestra aleatoria de la población al completo, sino que se eligen una colección de muestras aleatorias de tamaños prefijados de antemano de cada una de las r (o s) subpoblaciones que presentan un valor determinado de la característica X (Y).

La siguiente presentación tabular se denomina *tabla de contingencia*:

	Característica *Y*						
Característica *X*	**1**	**2**	**· · ·**	***j***	**· · ·**	***s***	**Total**
1	N_{11}	N_{12}	$\cdots$	N_{1j}	$\cdots$	N_{1s}	N_1
i	N_{i1}	N_{i2}	$\cdots$	N_{ij}	$\cdots$	N_{is}	N_i
r	N_{r1}	N_{r2}	$\cdots$	N_{rj}	$\cdots$	N_{rs}	N_r
Total	M_1	M_2	$\cdots$	M_j	$\cdots$	M_s	n

Problemas de repaso

1. Los siguientes datos representan los accidentes leves que han ocurrido durante el pasado año en una determinada planta industrial, clasificados de acuerdo con las horas del día en que tuvieron lugar.

Horas del día	**Número de accidentes**
08-10 h.	47
10–12 h.	52
13-15 h.	57
15-17 h.	63

Contraste la hipótesis de que los accidentes tienen la misma probabilidad de ocurrir en cualquiera de los cuatro periodos horarios incluidos en la tabla. Utilice un nivel de significación del 5%.

2. Una distribuidora cinematográfica exhibe preproyecciones no anunciadas de películas nuevas. En tal situación la película cuyo título no fue anunciado se proyecta además de la película programada. Cuando los espectadores dejan la sala se les entrega un cuestionario para que se rellene en casa y sea remitido por correo a la distribuidora. Ésta utiliza la información recibida para decidir en qué medida se ha de distribuir la película. Un tema que les interesa saber es si la aceptación de la película va a ser similar en las distintas partes del país. Para contrastar esta hipótesis, se han planeado preproyecciones no anunciadas de una película en cuatro salas de cine localizadas en ciudades distintas: Nueva York, Chicago, Phoenix y Seattle. Los siguientes datos muestran

los niveles de aceptación de la película por parte de los espectadores de las ciudades anteriores.

	Ciudad			
Aceptación	**Nueva York**	**Chicago**	**Phoenix**	**Seattle**
Excelente	234	141	108	142
Buena	303	256	165	170
Pobre	102	88	41	45

Al nivel de significación del 5%, contraste la hipótesis de que la aceptación de la película es independiente de la ciudad donde se proyectó. ¿Qué ocurre al nivel de significación del 1%?

3. Supongamos que se llevan a cabo l600 lanzamientos de un dado. Considere el contraste de la hipótesis de que las seis caras del dado son igualmente probables. Invente los datos que, en su opinión, producirían unos p valores aproximadamente iguales a

   (a) 0,50

   (b) 0,05

   (c) 0,95

   (d) Calcule los p valores reales para los datos propuestos en los apartados (a), (b) y (c).

4. Se intuye que, en el día de hoy, las proporciones de votantes a favor de cada uno de los candidatos demócrata, republicano e independiente en las futuras elecciones son del 40%, 42% y 18%, respectivamente. Para contrastar esta hipótesis se ha seleccionado una muestra aleatoria de 50 votantes; con ella se han obtenido los resultados siguientes:

	Demócrata	**Republicano**	**Independiente**
Número de personas a favor	18	22	10

Con los datos anteriores, ¿resulta consistente dicha intuición? Utilice un nivel de significación del 5%.

5. Los siguientes datos proceden de un análisis de accidentes de automóvil seleccionados aleatoriamente. Cada accidente se ha clasificado teniendo en cuenta el peso del coche accidentado y la severidad del daño sufrido por el conductor.

	Peso del coche (en libras)		
Daño	**Menos de 2500**	**2500-3000**	**Más de 3000**
Muy severo	34	22	8
Medio	43	41	47
Moderado	51	60	50

Al nivel de significación del 5%, contraste la hipótesis de que el daño del conductor y el peso del vehículo son independientes.

6. Un amigo nos indica que ha obtenido los siguientes resultados tras haber realizado 1000 lanzamientos de un dado:

Resultado	Frecuencia
1	167
2	165
3	167
4	166
5	167
6	168

¿Estos resultados son creíbles? ¡Explique por qué!

7. A partir de una muestra de 527 terremotos ocurridos en la parte occidental de Japón se han obtenido las siguientes frecuencias por los periodos horarios del día:

Resultado	Frecuencia
00-06 h	123
06-12 h	135
12-18 h	141
18-24 h	128

Contraste la hipótesis de que los terremotos tienen la misma probabilidad de ocurrir en cualquiera de los cuatro periodos horarios.

8. Los siguientes datos muestran el número de asesinatos, por día de la semana, cometidos en el estado de UTA entre 1978 y 1990:

Día	Domingo	Lunes	Martes	Miércoles	Jueves	Viernes	Sábado
Número	109	74	97	94	83	107	100

Contraste la hipótesis, al nivel de significación del 5%, de que un asesinato pudo haberse cometido en cualquier día de la semana con la misma probabilidad.

9. La tabla siguiente muestra la distribución porcentual acumulada de las alturas de las mujeres residentes en Estados Unidos que, entre 1976 y 1980, tuvieron una edad comprendida entre 18 y 24 años:

Altura	Porcentaje acumulado de mujeres
5 pies 0 pulgadas	4,22
5 pies 3 pulgadas	29,06
5 pies 5 pulgadas	58,09
5 pies 7 pulgadas	85,37
5 pies 8 pulgadas	92,30

Fuente: Centro Nacional de la Salud Pública de Estados Unidos, *Estadísticas de Vida y de Salud*, serie 11, n. 238.

La tabla indica, por ejemplo, que un 4,22% de las mujeres citadas tenían una altura menor o igual que 5 pies 0 pulgadas, un 24,84% tenían una altura mayor que 5 pies 0 pulgadas y menor o igual que 5 pies 3 pulgadas, y así sucesivamente.

Supongamos que se ha seleccionado una muestra aleatoria de 200 mujeres que tienen en la actualidad una edad comprendida ente 18 y 24 años; de éstas, 6 miden menos de 5 pies 0 pulgadas, 42 tienen una altura comprendida entre 5 pies 0 pulgadas y 5 pies 3 pulgadas, 48 de ellas están entre los 5 pies 3 pulgadas y 5 pies 5 pulgadas de altura, 60 miden entre 5 pies 5 pulgadas y 5 pies 7 pulgadas, 21 se encuentran entre 5 pies 7 pulgadas y 5 pies 8 pulgadas, y el resto tienen una altura superior a 5 pies 8 pulgadas. ¿Implican estos datos que la distribución de alturas ha cambiado? Utilice un nivel de significación del 5%.

10. En uno de los experimentos genéticos de Mendel resultaron los datos siguientes:

Tipo de guisante	Esperados	Observados
Amarillo liso	313	315
Amarillo rugoso	104	101
Verde liso	104	108
Verde rugoso	35	32

¿Se puede pensar que este ajuste es "demasiado bueno para que sea verdad"?

11. Una clínica de salud pública detalló los resultados relativos a 260 pacientes de edad avanzada a los que se les aconsejó que se vacunaran de la gripe. Un total de 184 de ellos efectivamente decidieron vacunarse, mientras que los restantes 76 no lo hicieron. Los resultados obtenidos para ambos grupos, referidos a los contagios de gripe durante la temporada, fueron los siguientes:

	Vacunados	No vacunados
Con gripe	10	6
Sin gripe	174	70

¿Prueban estos datos que existen diferencias en la probabilidad de contagio de la gripe entre los vacunados y los no vacunados? Utilice un nivel de significación del 5%. Si existieran diferencias, observe qué ocurre al nivel de significación del 1%; si no existieran diferencias, observe qué sucede al nivel de significación del 10%.

12. Se clasificó a una muestra aleatoria de 262 hombres casados con edades entre 50 y 60 años atendiendo al nivel de educación y al número de hijos de cada uno de ellos. Los datos obtenidos aparecen en la siguiente tabla de contingencia:

	Número de hijos		
Educación	**0–1**	**2–3**	**Más de 3**
Elemental	10	28	22
Secundaria	19	63	38
Universitaria	14	41	27

Contraste la hipótesis de que el tamaño familiar es independiente del nivel de educación del padre. Utilice un nivel de significación del 5%.

13. Los datos siguientes relacionan la edad de la madre con el peso (en gramos) de sus hijos recién nacidos:

	Peso del recién nacido (en gramos)	
Edad de la madre (en años)	**Menos de 2500**	**Más de 2500**
20 o menos	12	50
Más de 20	18	125

(a) Al nivel de significación del 5%, contraste la hipótesis de que los pesos de los bebés son independientes de las edades de las madres.

(b) ¿Cuál es el p valor resultante?

14. Repita el problema 13 tras haber multiplicado por 2 los cuatro valores de la tabla.

15. En una encuesta de evaluación docente, los estudiantes de un curso tenían que valorar el curso como excelente, regular o malo. Adicionalmente, los estudiantes tenían que indicar si el curso era para ellos obligatorio u opcional. Los datos obtenidos a partir de una muestra aleatoria de 121 estudiantes fueron los siguientes:

	Valoración del curso		
	Excelente	**Regular**	**Malo**
Obligatorio	14	42	18
Opcional	12	28	7

Al nivel de significación del 5%, contraste la hipótesis de que la valoración del curso es independiente de que sea obligatorio o no. ¿Qué ocurre al nivel del 1%?

16. Una clase de 154 estudiantes se imparte en un aula que tiene una capacidad de 250 personas. No por curiosidad, el profesor clasifica cada estudiante atendiendo a su sexo y la posición que ocupa en el aula. Utilice los siguientes datos para contrastar la hipótesis de que las dos clasificaciones son independientes.

	Parte delantera	**Parte central**	**Parte trasera**
Mujeres	22	40	18
Hombres	10	38	26

17. Se podría esperar que los primeros dígitos de los números contenidos en un anuario estadístico tienen la misma probabilidad de que sean 1, 2, . . . , o 9. Haga una selección aleatoria de números de un anuario estadístico y anote las primeras cifras de cada uno de ellos. Utilice estos datos para contrastar la hipótesis de que los 9 dígitos son igualmente probables.

18. Repita el problemas 17, pero utilice ahora la segunda cifra de cada número.

14

Contrastes de hipótesis no paramétricos

El hombre prefiere creer que lo que desea es cierto.

Francis Bacon (Moraleja: Uno debería contrastar sus propias hipótesis.)

Se consideran los contrastes de hipótesis en situaciones en las que la distribución de la población subyacente es desconocida y, además, no se puede asumir que dicha distribución pertenezca a una familia paramétrica dada, como, por ejemplo, la normal. Se ve cómo se puede utilizar el contraste de signos para contrastar hipótesis relativas a la mediana de una distribución. También se introduce el contraste de rangos signados, que permite contrastar la hipótesis de que una distribución poblacional es simétrica respecto de 0. Posteriormente, se presenta el contraste de la suma de rangos para contrastar la igualdad de dos distribuciones poblacionales. Finalmente se estudia el contraste de rachas que se puede utilizar para contrastar la hipótesis de que una sucesión de ceros y unos es una sucesión aleatoria que no sigue ningún patrón dado.

14.1 Introducción

¿Estamos propiciando el calentamiento del planeta? Más concretamente: ¿El comportamiento humano está causando un aumento en la temperatura de la Tierra? Incluso aunque los datos indiquen que las temperaturas anuales medias más recientes se sitúan entre las más altas registradas hasta ahora, esta pregunta tiene una muy difícil respuesta. Una dificultad afecta a los diferentes puntos geográficos donde se han tomado las medidas a lo

largo del tiempo. Por ejemplo, las antiguas mediciones de temperatura se llevaban a cabo en regiones rurales relativamente aisladas, mientras que en la actualidad las medidas se toman cerca de las ciudades, con un gran número de carreteras pavimentadas (que tienden a absorber calor). Este hecho, por sí mismo, redundaría en registros más altos de las temperaturas actuales. Otra dificultad proviene de la incertidumbre que afecta a la precisión de las medidas antiguas. Adicionalmente, se puede plantear el problema estadístico de si las temperaturas más altas actuales se deben a algún cambio real, tal como la combustión de productos derivados del carbón que podría producir el efecto invernadero al atrapar la energía del sol en la atmósfera de la tierra, o de si los actuales altos registros se deben a fluctuaciones debidas al azar que aparecen en las muestras aleatorias.

Para tener un punto de apoyo desde el que se pueda abordar la parte estadística de este problema sería interesante que fuéramos capaces de contrastar si el conjunto de datos sobre las temperaturas medias a lo largo del tiempo representa una muestra aleatoria proveniente de una determinada distribución de probabilidad, o si la propia distribución de temperaturas está cambiando a lo largo del tiempo.

Para poder abordar la cuestión de si existe una determinada distribución subyacente a las temperaturas observadas que se mantiene invariable a lo largo del tiempo es importante observar que no se está especificando por adelantado la forma de esta distribución. En particular, puesto que no existe a priori ninguna razón para creer que dicha distribución subyacente tenga por qué ser necesariamente una distribución normal, no se debería asumir esta hipótesis. Por el contrario, lo que querríamos es poder desarrollar un contraste de hipótesis que fuera válido para cualquier tipo de distribución subyacente. En este capítulo se estudiarán aquellos contrastes de hipótesis que se pueden usar en situaciones en las que no se requiere que la distribución subyacente a los datos tenga una forma particular. Dado que la validez de estos contrastes no se basa en la hipótesis previa de que la distribución subyacente pertenezca a una determinada familia paramétrica (tal como la normal) se dice que estos contrastes son *no paramétricos*.

14.2 Contraste de signos

Consideremos una población grande de elementos, cada uno de los cuales tiene asignado un determinado valor. Supongamos que la distribución de los valores poblacionales es continua y que estamos interesados en contrastar hipótesis que afectan a la mediana, como valor central, de esta distribución. Si la distribución de la población fuera normal, la mediana sería igual a la media; por consiguiente, se podrían emplear los contrastes vistos en los capítulos anteriores. Sin embargo, la normalidad aquí no se asumirá, sino que por el contrario se presentarán tipos de contrastes que se pueden utilizar para cualquier distribución subyacente continua.

Denotemos como η a la mediana de la población. Esto es, aquel valor que verifica que exactamente la mitad de los miembros de la población toman valores inferiores a η y la otra mitad toman valores superiores a él. Equivalentemente, si X es un miembro elegido aleatoriamente de la población, se cumple que

$$P\{X < \eta\} = P\{X > \eta\} = \frac{1}{2}$$

Supongamos ahora que se pretende contrastar la hipótesis nula de que la mediana es igual a un determinado valor m dado. Para poder contrastar

$$\mathrm{H}_0\colon \eta = m$$

frente a

$$H_1: \eta \neq m$$

denotemos por p la proporción de miembros de la población cuyos valores son menores que m. Esto es,

$$p = P\{X < m\}$$

siendo X un miembro aleatoriamente elegido de la población. Ahora bien, si la hipótesis nula es cierta y, en consecuencia, m es realmente la mediana, el valor de p será igual a 1/2. Por el contrario, si m no coincide con la mediana, el valor de p diferirá de 1/2. En consecuencia, el contraste de la hipótesis de que la mediana es igual a m es equivalente al contraste de la hipótesis nula

$$H_0: p = \frac{1}{2}$$

frente a la alternativa

$$H_1: p \neq \frac{1}{2}$$

Así pues, se ve que contrastar la hipótesis de que la mediana es igual a m equivale a contrastar si una determinada proporción poblacional es igual a 1/2. Esta proporción coincide, naturalmente, con la proporción de individuos de la población con valores menores que m.

Se pueden utilizar los resultados de la sección 9.5.1 para contrastar la hipótesis nula de que la mediana poblacional es igual a m. Concretamente, seleccionemos una muestra de n elementos de la población y denotemos por TS el número de ellos que tienen valores inferiores a m. Observe que, si H_0 es cierta, TS es una variable aleatoria binomial de parámetros n y 1/2. El contraste consistirá en rechazar la hipótesis nula si el valor de TS resulta ser demasiado grande o demasiado pequeño. Específicamente, si el valor observado de TS es i, el contraste a nivel de significación α rechazará H_0 si

$$P\{N \geq i\} \leq \frac{\alpha}{2}$$

o bien si

$$P\{N \leq i\} \leq \frac{\alpha}{2}$$

siendo N una variable aleatoria binomial con parámetros $(n, 1/2)$. La figura 14.1 ilustra la forma de actuar del contraste.

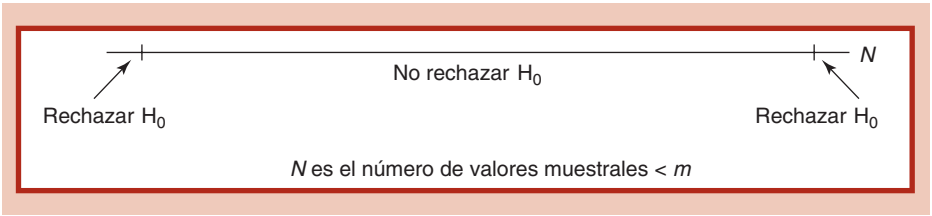


Figura 14.1 Contraste de H_0: $\eta = m$ frente a H_1: $\eta \neq m$.

Puesto que se ha asumido que la distribución es continua, en principio no debería haber ningún valor de dato igual a m. Pese a ello, dado que las mediciones se registran de acuerdo con la precisión de los aparatos utilizados, lo anterior sí que podría ocurrir en la práctica. Si existieran valores muestrales iguales a m, éstos se deberían eliminar de la muestra y, en consecuencia, reducir el valor de n.

En términos del p valor, lo anterior se puede resumir como sigue.

Para contrastar

$$H_0: \eta = m \qquad \text{frente a} \qquad H_1: \eta \neq m$$

seleccione una muestra aleatoria y suprima los valores iguales a m. Sea n el número de valores muestrales no suprimidos. Definamos como estadístico del contraste el número de tales valores muestrales inferiores a m. Si dicho número es i, el p valor coincide con

$$p \text{ valor} = 2 \text{ Min}(P\{N \leq i\}, P\{N \geq i\})$$

donde N representa una variable aleatoria binomial con parámetros n y 1/2. Así, la hipótesis nula se rechazará a cualquier nivel de significación mayor o igual que el p valor, y no se rechazará en otro caso.

Para encontrar el p valor no es necesario obtener las dos probabilidades $P\{N \leq i\}$ y $P\{N \geq i\}$. Tan sólo se ha de calcular el mínimo de ambas. Ahora bien, puesto que $E[N] = n/2$, el mínimo coincidirá con $P\{N \leq i\}$ cuando i sea pequeño (comparado con $n/2$), y coincidirá con $P\{N \geq i\}$ cuando i sea grande (comparado con $n/2$). Cuando i esté próximo a $n/2$ no será necesario cálculo alguno, puesto que ambas probabilidades se aproximarán a 1/2 y que el p valor estará muy próximo a 1. Así pues, desde un punto de vista práctico, el p valor se puede expresar como

$$p \text{ valor} = \begin{cases} 2P\{N \leq i\} & \text{si } i \leq \dfrac{n}{2} \\ 2P\{N \geq i\} & \text{si } i \geq \dfrac{n}{2} \end{cases}$$

donde N es una binomial con parámetros n y 1/2.

Ejemplo 14.1 La política de inventario de una zapatería se basa en la creencia de que la mediana de las longitudes de pie de los adolescentes es igual a 10,25 pulgadas. Para contrastar esta hipótesis se midieron los pies de los 50 miembros de una muestra aleatoria de adolescentes. Supongamos que 36 de los citados adolescentes tienen unas longitudes de pie superiores a 10,25 pulgadas. ¿Está esto en contra de la hipótesis de que la talla mediana de los pies es de 10,25?

Solución Sea N una variable aleatoria binomial con parámetros (50, 1/2). Dado que 36 es mayor que $50(1/2) = 25$, se tiene que el p valor es

$$p \text{ valor} = 2P\{N \geq 36\}$$

Se puede utilizar la aproximación normal, o el Programa 5-1, para calcular esta probabilidad. Puesto que

$$E[N] = 50 \times \frac{1}{2} = 25 \qquad \text{Var}(N) = 50 \times \frac{1}{2} \times \frac{1}{2} = 12{,}5$$

la aproximación normal proporciona los valores siguientes:

$$\begin{aligned} p \text{ valor} &= 2P\{N \geq 36\} \\ &= 2P\{N \geq 35{,}5\} \quad \text{(debido a la corrección por continuidad)} \\ &= 2P\left\{\frac{N - 25}{\sqrt{12{,}5}} \geq \frac{35{,}5 - 25}{\sqrt{12{,}5}}\right\} \\ &\approx 2P\{Z \geq 2{,}97\} \\ &= 0{,}0030 \quad \text{a partir de la tabla D.1} \end{aligned}$$

(El Programa 5-1, que computa las probabilidades de la binomial, obtiene como p valor exacto 0,0026.) Así pues, la creencia de que la talla mediana de pies es igual a 10,25 pulgadas debe rechazarse incluso al nivel de significación del 1%. Parece, pues, existir una fuerte evidencia de que la talla mediana es mayor que 10,25. ■

Supongamos que $X_1, \ldots, X_n$ representan los n valores muestrales. Dado que el valor del estadístico del contraste depende solamente de los signos, positivos o negativos, de los valores $X_i - m$, el contraste anterior se conoce con el nombre de *contraste de signos*.

14.2.1 Contraste de igualdad de distribuciones poblacionales cuando las muestras son apareadas

El contraste de signos se puede utilizar para comparar dos poblaciones cuando existe un apareamiento natural de los elementos de las muestras. Se ilustrará esto con un ejemplo.

Ejemplo 14.2 Se ha llevado a cabo un experimento para ver si dos cremas de protección solar de factor 15 son igualmente efectivas. Un grupo de 12 voluntarios tomaron el sol boca abajo al mediodía durante una hora. A cada voluntario se le aplicó la crema solar A en un lado de su espalda y la crema B en el otro. Posteriormente, para cada voluntario, se midieron las radiaciones solares recibidas en ambos lados. Si 10 de los voluntarios tenían menos quemado el lado donde se aplicó la crema A que el lado donde se aplicó la crema B, ¿se puede concluir que las dos cremas solares no tienen igual efectividad?

Solución Se puede concebir que se está ante dos poblaciones diferentes, la población de los lados de espalda donde se aplicó el protector solar A y la población de los lados donde se aplicó el protector B. Los elementos "apareados" de las dos muestras son los dos lados correspondientes a un mismo voluntario. Si las dos cremas solares fueran igualmente efectivas, la mediana de la diferencia entre las quemaduras solares de los dos lados de la espalda de cada voluntario sería igual a 0. Esto es, por mero azar, aproximadamente la mitad de los voluntarios reflejarían que la crema A se comportó mejor que la B, y viceversa. Así pues,

se puede contrastar que la efectividad de las dos cremas es igual sin más que contrastar la hipótesis de que la mediana de las diferencias entre las quemaduras recibidas por cada voluntario con las cremas A y B es igual a 0.

Puesto que el número de diferencias con valor negativo es 10, que es mayor que $12(1/2) = 6$, se obtiene a partir del contraste de signos que el p valor coincide con

$$p \text{ valor} = 2P\{N \geq 10\}$$

donde N es una binomial de parámetros (12, 1/2). Puesto que

$$\begin{aligned} P\{N \geq 10\} &= P\{N = 10\} + P\{N = 11\} + P\{N = 12\} \\ &= \frac{12!}{10!\,2!}\left(\frac{1}{2}\right)^{12} + \frac{12!}{11!\,1!}\left(\frac{1}{2}\right)^{12} + \frac{12!}{12!\,0!}\left(\frac{1}{2}\right)^{12} \\ &= \left[\frac{12 \cdot 11}{2 \cdot 1} + 12 + 1\right]\left(\frac{1}{2}\right)^{12} = \frac{79}{4096} \end{aligned}$$

se ve que

$$p \text{ valor} = \frac{158}{4096} = 0{,}0386$$

Así pues, se debe rechazar la hipótesis de que los dos protectores solares son igualmente efectivos a cualquier nivel de significación mayor o igual que 3,86%. (Por ejemplo, se deberá rechazar al nivel de significación del 5%, pero no al nivel de significación del 1%.) ■

14.2.2 Contrastes unilaterales

También se puede utilizar el contraste de signos para llevar a cabo contrastes de hipótesis unilaterales con respecto a la mediana. Supongamos que se desea contrastar

$$H_0: \eta \leq m$$

frente a

$$H_1: \eta > m$$

donde η es la mediana poblacional y m un valor cualquiera prefijado de antemano. De nuevo, denotemos por p la proporción de elementos de la población cuyos valores son menores que m. Ahora bien, si la hipótesis nula es cierta y, en consecuencia, m es al menos tan grande como η, la proporción de elementos de la población cuyo valor es menor que m es como mínimo 1/2 (véase la figura 14.2). Similarmente, si la hipótesis alternativa es cierta y, por consiguiente, m es menor que η, la proporción de elementos de la población con valores inferiores a m es menor que 1/2 (véase de nuevo la figura 14.2). De aquí se deduce que el contraste anterior equivale a contrastar

$$H_0: p \geq \frac{1}{2}$$

frente a

$$H_1: p < \frac{1}{2}$$

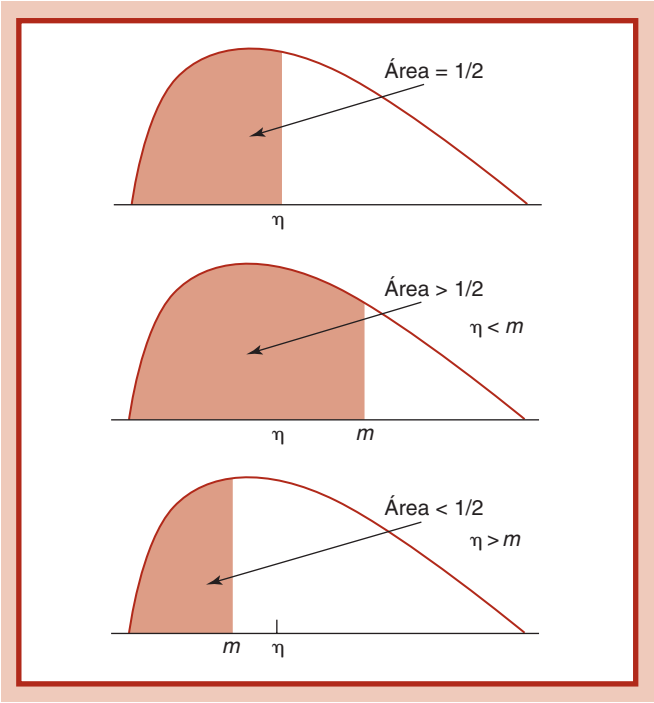


Figura 14.2 $P\{X < m\} < 1/2$ si $\eta > m$ $P\{X < m\} = 1/2$ si $\eta = m$ $P\{X < m\} > 1/2$ si $\eta < m$.

Para utilizar el contraste de signos cuando se pretende contrastar la anterior hipótesis nula unilateral se selecciona una muestra aleatoria de n elementos de la población. Supongamos que i de ellos presentan valores inferiores a m. El p valor resultante coincide con la probabilidad de obtener un valor menor o igual que i, si cada elemento de la población tuviera una probabilidad 1/2 de tomar un valor inferior o igual a m. Esto es, si N representa una variable aleatoria binomial de parámetros $(n, 1/2)$, se tiene que

$$p \text{ valor} = P\{N \leq i\}$$

Ejemplo 14.3 Un banco ha decidido abrir una nueva oficina en una determinada zona si la renta anual mediana de sus residentes supera los 40 000 dólares. Para obtener esta información, se seleccionó una muestra aleatoria de 80 familias, que fueron encuestadas acerca de sus rentas. De las 80 familias, 52 tenían rentas anuales superiores al umbral de 40 000 dólares fijado por el banco, y 28 las tenían inferiores. ¿Es esta información lo suficientemente significativa, al nivel de significación del 5%, para establecer que la renta mediana de las familias residentes en la zona es mayor que 40 000 dólares?

Solución Se debe averiguar si los datos tienen la fuerza suficiente como para rechazar la hipótesis nula cuando se contrasta

$$H_0: \eta \leq 40 \qquad \text{frente a} \qquad H_1: \eta > 40$$

Si p es la proporción de familias de la población con rentas anuales inferiores a 40 000 dólares, lo anterior equivale a contrastar

$$H_0: p \geq \frac{1}{2} \qquad \text{frente a} \qquad H_1: p < \frac{1}{2}$$

Puesto que 28 de las 80 familias muestreadas tienen rentas anuales inferiores a 40 000 dólares, el p valor de estos datos es

$$p \text{ valor} = P\{N \leq 28\}$$

donde N representa una variable aleatoria binomial de parámetros (80, 1/2). Si se utiliza el Programa 5-1 (o la aproximación a la normal) se obtiene

$$p \text{ valor} = 0{,}0048$$

Para este p valor tan pequeño se debe rechazar la hipótesis nula de que la renta mediana es inferior o igual a 40 000 dólares y se concluirá, pues, que casi con certeza sobrepasa el umbral fijado por el banco. ■

El contraste de la hipótesis nula de que la medina es mayor o igual a m es similar al anterior. Así pues, los contrastes con una hipótesis nula unilateral se sintetizan como sigue.

Contraste de la mediana con una hipótesis nula unilateral

Para contrastar bien

$$H_0: \eta \leq m \qquad \text{frente a} \qquad H_1: \eta > m \tag{1}$$

o

$$H_0: \eta \geq m \qquad \text{frente a} \qquad H_1: \eta < m \tag{2}$$

seleccione una muestra aleatoria de elementos de la población. Elimine aquellos que tienen valores iguales a m y denote como n al número de elementos muestrales no eliminados. Sea TS el número de valores de datos menores que m. Si TS es igual a i, los p valores son

$$p \text{ valor} = P\{N \leq i\} \quad \text{en el caso (1)}$$
$$p \text{ valor} = P\{N \geq i\} \quad \text{en el caso (2)}$$

donde N representa una variable aleatoria binomial con parámetros n y 1/2.

Problemas

1. La cifra publicada como mediana de la presión sanguínea sistólica de los hombres de edad intermedia es 128. Para determinar si se ha producido un cambio en este valor se ha seleccionado una muestra aleatoria de 100 hombres de dicha edad. Contraste la hipótesis de que la mediana es igual a 128 si:

   (a) 60 hombres de la muestra presentaban presiones superiores a 128.

   (b) Fueron 70 los que tenían presiones por encima de 128.

   (c) 80 hombres presentaron presiones superiores a 128.

   En cada caso, calcule el p valor.

2. En el año 2001 la renta mediana de los hogares del estado de Connecticut era de 52 758 dólares. En una encuesta reciente se muestrearon aleatoriamente 250 hogares y se descubrió que el 42% de ellos tenía rentas inferiores a la mediana de 2001 y que el 58% restante tenía rentas por encima de ésta. A partir de estos datos ¿se puede establecer que la renta mediana de los hogares de Connecticut ha dejado de ser igual a la de 2001? ¿Cuál es el p valor?

3. Cincuenta estudiantes de una academia de policía han hecho prácticas de tiro, utilizando dos tipos de pistolas. Cada estudiante realizó la mitad de los disparos con una pistola barata, y la otra mitad con una pistola más cara. Si 29 de los estudiantes consiguieron puntuaciones más elevadas con la pistola barata, ¿permite esto establecer que las dos pistolas no son igualmente efectivas? Utilice un nivel de significación del 5%.

4. Una clínica de dermatología pretende comparar la efectividad de una nueva crema de manos con la de la crema que actualmente recomienda a los pacientes que presentan un determinado tipo de eczema. Para recoger información se pidió a la mitad de sus pacientes que se aplicaran la crema nueva en la mano izquierda y la crema antigua en la derecha, cada noche durante una semana; mientras que a la otra mitad de los pacientes se les indicó que se extiendan cada noche la crema nueva en la mano derecha y la antigua en la izquierda durante el mismo periodo. Tras ello, todos los pacientes fueron examinados. Supongamos que, para el 60% de los pacientes, se mostraba que la mano que había sido tratada con la crema nueva había mejorado más que la mano tratada con la crema antigua.

   Si el número de pacientes que tomaron parte en el experimento fue igual a:

   (a) 10 (b) 20 (c) 50 (d) 100 (e) 500

   ¿prueban estos datos que las dos cremas no son igualmente efectivas? Utilice un nivel de significación del 5%. Adicionalmente, calcule el p valor resultante en cada caso.

5. Una profesora de Estadística ha ideado una propuesta de examen para una clase con un gran número de estudiantes. Pretende que la calificación mediana del examen sea al menos de 72 puntos y piensa que su propuesta se adecua a ese objetivo. Para ser cuidadosa, eligió aleatoriamente a 13 estudiantes para que hicieran el examen previamente al resto. Si sus calificaciones fueron

   $$65, 79, 77, 90, 56, 60, 65, 80, 70, 69, 83, 69, 65$$

   ¿debería rechazar la hipótesis de que la mediana de las calificaciones será al menos de 72 puntos con ese examen? Utilice un nivel de significación del 5%.

6. La mediana de los precios de venta de las casas de una determinada zona residencial se ha mantenido constante e igual a 122 000 dólares durante los 2 últimos años. Para averiguar si el precio mediano ha aumentado se ha seleccionado una muestra de 20 casas vendidas recientemente. Sus precios de venta fueron los siguientes (en unidades de 1000 $):

   $$144, 116, 125, 128, 96, 92, 163, 130, 120, 142, 155,$$
   $$133, 110, 105, 136, 140, 124, 130, 88, 146$$

   ¿Son estos datos lo suficientemente convincentes para establecer que el precio mediano ha aumentado? Utilice un nivel de significación del 5%.

7. Para contrastar la hipótesis de que el peso mediano de las chicas de 16 años de la ciudad de Los Ángeles es como mínimo de 110 libras se seleccionó una muestra de 200 chicas.

Si 120 de ellas pesaron menos de 110 libras, ¿está esto en contra de la hipótesis? Utilice un nivel de significación del 5%. ¿Cuál es el p valor resultante?

8. Para intentar probar que el aceite de pescado reduce los niveles sanguíneos de colesterol, un médico nutricionista pidió a 24 voluntarios que añadieran a su dieta cierto aceite de pescado durante 3 meses. Tras ese periodo chequeó los niveles de colesterol de todos ellos y observó que, en 16 de los 24 voluntarios, sus niveles de colesterol se habían reducido tras la prueba.

   (a) ¿Cuáles son las hipótesis nula y alternativa?

   (b) ¿Se debe rechazar la hipótesis nula? Utilice un nivel de significación del 5%.

   (c) ¿Cuál es el p valor resultante?

14.3 Contraste de rangos signados

En la sección 14.2.1 se vio cómo se puede utilizar el contraste de signos para contrastar la hipótesis nula de que dos poblaciones tengan la misma distribución de valores, cuando los datos están formados por muestras apareadas. Para contrastar esta hipótesis se consideraron las diferencias entre los valores muestrales apareados. Se tuvo en cuenta que, cuando la hipótesis nula es cierta, la mediana de estas diferencias es 0, y el contraste de signos se utilizaba, pues, para contrastar esta última hipótesis.

La única información necesaria para el contraste de signos sobre la igualdad de dos distribuciones poblacionales, cuando se utilizan muestras apareadas, es el número de veces que el primer dato del par es mayor que el segundo. Esto es, el contraste de signos no requiere que se conozcan los valores reales de los pares de datos, sino que sólo precisa conocer cuál de ellos es el mayor. Sin embargo, aunque su aplicación es sencilla, el contraste de signos no es particularmente eficiente para contrastar la hipótesis nula de que las dos distribuciones poblacionales coincidan. Esto se debe a que, si la hipótesis nula es cierta, no únicamente la distribución de la diferencia de los valores apareados tendrá mediana cero, sino que también verificará una propiedad más fuerte, que es la de ser simétrica respecto de cero. Es decir, para cualquier valor x es igualmente probable que el primer valor del par supere al segundo en la cuantía x como que el segundo valor del par supere al primero en esa misma cuantía (véase la figura 14.3). El contraste de signos no tiene en cuenta la simetría de la distribución de las diferencias, sino únicamente que la mediana de esta distribución sea cero.

Por ejemplo, supongamos que los datos se componen de 12 valores apareados cuyas diferencias son las siguientes:

$$2,\ 5,\ -0{,}1,\ -0{,}4,\ -0{,}3,\ 9,\ 7,\ 8,\ 12,\ -0{,}5,\ -1,\ -0{,}6$$

Puesto que seis de las diferencias son positivas y seis son negativas, este conjunto de datos es perfectamente consistente con la hipótesis de que la mediana de las diferencias sea 0. Sin embargo, por otro lado, ya que los valores altos son positivos, los datos no aparentan ser consistentes con la hipótesis de que la distribución es simétrica respecto de 0. Así pues, parece altamente improbable que las distribuciones poblacionales sean iguales.

Supongamos de nuevo que se desean utilizar los datos provenientes de muestras apareadas para contrastar la hipótesis de que las dos distribuciones poblacionales son iguales. Se

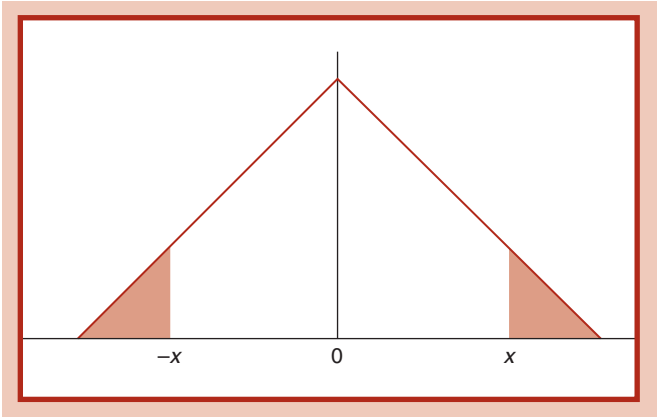


Figura 14.3 Densidad simétrica respecto a 0. Las zonas sombreadas de las áreas son iguales.

presentará a continuación un contraste más sensible que el contraste de signos. Se conoce con el nombre de *contraste de rangos signados*, y actúa contrastando si la distribución de las diferencias de los valores apareados es simétrica respecto de 0.

Supongamos que se han seleccionado las muestras apareadas de tamaño n de las dos poblaciones. Denotemos como D_i a la diferencia entre el valor de la primera población y el de la segunda correspondiente al par i-ésimo, para $i = 1, \ldots, n$. Ordenemos a continuación estas diferencias de acuerdo con su valor absoluto. Es decir, la primera diferencia ordenada debería ser aquel D_i con menor valor absoluto, y así sucesivamente. El estadístico del contraste de rangos signados es la suma de los rangos de (o sea, las posiciones que ocupan) los valores negativos una vez ordenados con el criterio anterior.

Ejemplo 14.4 Supongamos que los datos consisten en los siguientes cuatro valores muestrales apareados.

i	X_i	Y_i
1	4,6	6,2
2	3,8	1,5
3	6,6	11,7
4	6,0	2,1

Las diferencias $X_i - Y_i$ son, pues,

$$-1{,}6,\ 2{,}3,\ -5{,}1,\ 3{,}9$$

Si se ordenan estas diferencias de menor a mayor de acuerdo con su valor absoluto resulta lo siguiente:

$$-1{,}6,\ 2{,}3,\ 3{,}9,\ -5{,}1$$

Puesto que los rangos de los valores negativos son 1 y 4, resulta que el valor del estadístico del contraste de rangos signados es

$$\text{TS} = 1 + 4 = 5$$

En otras palabras, puesto que las diferencias negativas ocupan las posiciones ordenadas 1 y 4, TS = 1 + 4 = 5. ■

El contraste de rangos signados coincide con el contraste de signos en que ambos tienen en cuenta aquellos pares de datos para los que el valor de la primera población es menor que el de la segunda. Sin embargo, mientras que el contraste de signos da igual peso a cada uno de los pares, el contraste de rangos signados da mayor peso a aquellos pares cuyas diferencias están más alejadas de cero.

El contraste de rangos signados para contrastar

$$H_0\text{: las dos distribuciones poblacionales son iguales}$$

frente a

$$H_1\text{: las dos distribuciones poblacionales no son iguales}$$

propone rechazar la hipótesis nula si el estadístico del contraste TS es suficientemente grande o suficientemente pequeño. Un valor alto de TS indica que la mayoría de los valores altos de las diferencias tienen signo negativo; mientras que un valor pequeño de TS indica que la mayoría de los valores altos de esas diferencias tienen signo positivo. Cualquiera de estas situaciones supondría una prueba en contra de la simetría de la distribución de las diferencias y, en consecuencia, en contra de H_0.

Si el valor del estadístico del contraste es t, el contraste de rangos signados rechaza H_0 bien si

$$P\{\text{TS} \leq t\} \leq \frac{\alpha}{2}$$

o bien si

$$P\{\text{TS} \geq t\} \leq \frac{\alpha}{2}$$

Arriba, α es el nivel de significación, y las probabilidades se calcularán asumiendo que H_0 es cierta. Equivalentemente, el p valor resultante se calcula como sigue:

Supongamos que el valor de TS es t. El p valor del contraste de rangos signados viene dado por

$$p \text{ valor} = 2\,\text{Min}(P\{\text{TS} \leq t\}, P\{\text{TS} \geq t\})$$

donde las probabilidades se deben calcular asumiendo que H_0 es cierta.

Para llevar a cabo el contraste de rangos signados es necesario poder calcular las probabilidades anteriormente indicadas. El punto clave para conseguir esto se basa en el hecho de que

cuando H_0 es cierta y, en consecuencia, la distribución de las diferencias es simétrica respecto de cero, cada una de las diferencias, independientemente de las restantes, tiene la misma probabilidad de ser positiva o negativa. El Programa 14-1 utiliza este hecho para obtener explícitamente las probabilidades necesarias y el *p* valor resultante. Las entradas que se deben proporcionar al programa son el tamaño muestral *n* y el valor del estadístico del contraste TS.

Ejemplo 14.5 Un profesor de Psicología quiere saber si sus estudiantes obtendrían calificaciones similares en dos exámenes distintos. Para ello, seleccionó a 12 estudiantes que aceptaron participar en un experimento que diseñó al efecto. Seis de los estudiantes se sometieron al examen A, y los otros seis al B. Al día siguiente, los estudiantes realizaron el examen que no habían hecho el día anterior. Así pues, cada uno de los 12 estudiantes realizó los dos exámenes. A continuación se muestran los pares de calificaciones obtenidas por los estudiantes en los dos exámenes:

	Estudiante											
Examen	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
A	763	419	586	920	881	758	262	332	717	909	940	835
B	797	404	576	855	762	707	195	341	728	817	947	849

Así, por ejemplo, el estudiante 3 obtuvo 586 puntos en el examen A y 576 puntos en el B. Las diferencias de los valores apareados son las siguientes:

$$-34, 15, 10, 65, 119, 51, 67, -9, -11, 92, -7, -14$$

Si se ordenan estas diferencias en orden creciente de sus valores absolutos se obtiene

$$-7, -9, 10, -11, -14, 15, -34, 51, 65, 67, 92, 119$$

Puesto que las diferencias negativas se encuentran en las posiciones 1, 2, 4, 5 y 7, el valor del estadístico del contraste es

$$\text{TS} = 1 + 2 + 4 + 5 + 7 = 19$$

Para obtener el *p* valor, ejecutemos el Programa 14-1, que computa el *p* valor correspondiente al contraste de rangos signados para evaluar si la distribución poblacional es simétrica respecto de cero. El tamaño muestral es 12, y el valor observado de la suma de rangos signados es 19. El *p* valor que nos proporciona el Programa 14-1 es 0,1293945.

Puesto que el *p* valor es 0,129, no se puede rechazar la hipótesis nula de que las distribuciones de las calificaciones obtenidas por los estudiantes en los exámenes sean idénticas, al nivel de significación del 10%. ■

Si se tiene en cuenta el hecho de que las diferencias ordenadas son variables aleatorias independientes con la misma probabilidad de ser positivas o negativas, se puede demostrar que, cuando H_0 es cierta, la media y la varianza de TS vienen, respectivamente, dadas por

$$E[\text{TS}] = \frac{n(n+1)}{4}$$

y

$$\text{Var(TS)} = \frac{n(n+1)(2n+1)}{24}$$

Adicionalmente, para valores moderadamente grandes de n, se puede demostrar igualmente que, si H_0 es cierta, TS sigue una distribución que es aproximadamente normal con media y varianza iguales a las anteriormente indicadas. Este hecho nos permite aproximar el p valor si no disponemos del Programa 14-1.

Ejemplo 14.6 Veamos la precisión de la anterior aproximación por normalidad del p valor cuando se utilizan los datos del ejemplo 14.5. Puesto que $n = 12$, de las expresiones anteriores se obtiene que, cuando H_0 es cierta,

$$E[\text{TS}] = \frac{12 \cdot 13}{4} = 39 \qquad \text{Var(TS)} = \frac{12 \cdot 13 \cdot 25}{24} = 162{,}5$$

El valor del estadístico del contraste es 19. Puesto que este valor es menor que $E[\text{TS}]$, está claro que $P\{\text{TS} \leq 19\}$ es menor que $P\{\text{TS} \geq 19\}$. Por consiguiente,

$$\begin{aligned} p \text{ valor} &= 2P\{\text{TS} \leq 19\} \\ &= 2P\{\text{TS} \leq 19{,}5\} \quad \text{(corrección por continuidad)} \\ &= 2P\left\{\frac{\text{TS} - 39}{\sqrt{162{,}5}} \leq \frac{19{,}5 - 39}{\sqrt{162{,}5}}\right\} \\ &\approx 2P\{Z \leq -1{,}530\} \\ &= 0{,}126 \end{aligned}$$

Así pues, la aproximación normal obtiene un p valor que está bastante próximo a 0,129, el p valor real obtenido en el ejemplo 14.5. ■

Una regla empírica es que la aproximación normal del p valor es buena siempre que n, el número de datos apareados, sea como mínimo 25. (De hecho, este número es bastante conservador, puesto que en el ejemplo anterior la aproximación resultó ser buena pese a que n era igual a 12.)

Ejemplo 14.7 Supongamos que se ha obtenido un valor de TS igual a 238 con una muestra de 25 datos apareados. Si se asume que la distribución de las diferencias es simétrica respecto de 0, se obtiene, a partir de las fórmulas indicadas anteriormente, que

$$E[\text{TS}] = \frac{25 \cdot 26}{4} = 162{,}5$$

$$\sqrt{\text{Var(TS)}} = \sqrt{\frac{25 \cdot 26 \cdot 51}{24}} = 37{,}165$$

La aproximación normal del p valor es, pues,

$$\begin{aligned} p \text{ valor} &= 2P\{\text{TS} \geq 238\} \\ &= 2P\{\text{TS} \geq 237{,}5\} \\ &= 2P\left\{\frac{\text{TS} - 162{,}5}{37{,}165} \geq \frac{237{,}5 - 162{,}5}{37{,}165}\right\} \\ &\approx 2P\{Z \geq 2{,}018\} \\ &= 0{,}0436 \end{aligned}$$

Por otro lado, si se utiliza el Programa 14-1 se obtiene como p valor exacto:

$$p \text{ valor} = 0{,}0422$$

Así pues, se ve de nuevo que la aproximación está bastante cercana al valor exacto. ■

14.3.1 Diferencias nulas y empates

Si una diferencia tiene el valor 0 (debido a que los valores apareados son iguales), el par de datos se tendría que eliminar y el valor de n se reduciría en una unidad.

Si algunas de las diferencias tienen el mismo valor absoluto, el peso dado a una diferencia negativa debería ser el promedio de los rangos de todas las diferencias con el mismo valor absoluto. Por ejemplo, si las diferencias fueran

$$-1, 3, 8, -3$$

las diferencias ordenadas serían

$$-1, 3, -3, 8$$

Puesto que existe un empate entre las diferencias segunda y tercera, el valor del estadístico del contraste es

$$\text{TS} = 1 + \frac{2 + 3}{2} = 3{,}5$$

El Programa 14-1 no se debe utilizar si se producen empates. En su lugar, se empleará la aproximación normal indicada anteriormente.

La tabla 14.1 sintetiza todo lo relacionado con el contraste de rangos signados.

Problemas

1. Obtenga el valor del estadístico del contraste de rangos signados si las diferencias de los valores apareados son las siguientes:

   (a) $-17, 33, 22, -8, 55, -41, -18, 40, 39, 14, -88, 99, 102, -5, 7$

Tabla 14.1 Contraste de rangos signados

Permite contrastar la hipótesis de que dos distribuciones poblacionales son iguales cuando se dispone de muestras apareadas, siendo $X_1, \ldots, X_n$ los valores de la muestra procedente de la primera población y siendo $Y_1, \ldots, Y_n$ los valores de la muestra extraída de la segunda. Para $i = 1, \ldots, n$, los datos X_i e Y_i están apareados y su diferencia se denotará por $D_i = X_i - Y_i$.

Para contrastar

H_0: la distribución de las diferencias D_i es simétrica respecto de 0

frente a

H_1: la distribución anterior no es simétrica respecto de 0

elimine primero aquellos D_i iguales a 0 y rebaje el valor de n de forma que éste represente el número de las diferencias no nulas. Haga TS igual a la suma de las posiciones, o rangos, de las diferencias D_i negativas, una vez que éstas se hayan ordenado en valor absoluto de menor a mayor. Si dos o más de los valores absolutos de los D_i son iguales, se les dará a todos ellos un rango (o posición) igual al promedio de sus rangos.

Si TS $= t$, el p valor viene dado por

$$p \text{ valor} = 2 \text{ Min}(P\{\text{TS} \leq t\}, P\{\text{TS} \geq t\})$$

donde las probabilidades se han de calcular asumiendo que H_0 es cierta. Éstas se pueden aproximar teniendo en cuenta que, si H_0 es cierta, TS es aproximadamente una variable aleatoria normal con media y varianza dadas, respectivamente, por

$$E[\text{TS}] = \frac{n(n+1)}{4} \qquad \text{Var(TS)} = \frac{n(n+1)(2n+1)}{24}$$

Si no existen empates (es decir, diferencias iguales en valor absoluto), el p valor exacto se puede obtener con el Programa 14-1.

(b) 44, 2, 1, −0,4, −3, −13, 44, 50, 1,1, −2,2, 0,01, −4, −6,6

(c) 12, 15, 19, 8, −3, −7, −22, −55, 48, 31, 89, 92

2. Asumiendo que las diferencias entre los valores apareados siguen una distribución que es simétrica respecto de 0, calcule la media y la varianza del estadístico del contraste de rangos signados para cada uno de los apartados del problema 1.

3. Para cada apartado del problema 1 encuentre el p valor para el contraste de la hipótesis de que la distribución de las diferencias es simétrica respecto de 0. Utilice la aproximación normal.

4. Compare los resultados obtenidos en el problema 3 con los p valores exactos proporcionados por el Programa 14-1.

5. Una profesora de Historia se cuestiona si las calificaciones que obtienen sus alumnos en sus trabajos de casa dependen de si están escritos a mano o con un procesador de textos. Para averiguarlo diseñó un experimento en el que un grupo de 28 estudiantes fue dividido en 14 parejas, cuyos miembros tenían, en opinión de la profesora, aproximadamente el mismo nivel. Después asignó un trabajo y pidió que un miembro de cada pareja le presentara el trabajo escrito a mano y el otro miembro lo presentara procesado con ordena-

dor. Seleccionó al azar, lanzando una moneda, qué estudiante de cada pareja presentaría el trabajo escrito a mano. Las calificaciones dadas a los trabajos fueron las siguientes:

Pareja	Escrito a mano	Escrito con procesador
1	83	88
2	75	91
3	75	72
4	60	70
5	72	80
6	55	65
7	94	90
8	85	89
9	78	85
10	96	93
11	80	86
12	75	79
13	66	64
14	55	68

(a) ¿Se debe concluir que el hecho de que el trabajo se presente escrito a mano o con procesador influye en la calificación obtenida? Utilice un nivel de significación del 5%.

(b) ¿Cuál es el p valor resultante?

6. Para contrastar si un protector dental era efectivo para reducir las caries bucales se trataron con él la mitad de los dientes de 100 niños, mientras que los de la otra mitad no se trataron. Al cabo de 6 meses se observó la diferencia entre el número de caries de la parte tratada y el de la no tratada. El estadístico del contrate de rangos signados fue 1830. Al nivel de significación del 5%, ¿se puede concluir que el tratamiento con el protector establece diferencias distribucionales en el número de caries? ¿Qué ocurre al nivel de significación del 1%?

7. Una organización de consumidores desea determinar si los talleres de reparación de automóviles dan presupuestos distintos a las mujeres y los hombres. Para ello seleccionó dos coches con idénticas averías que fueron entregados, respectivamente, a un hombre y a una mujer. Después de elegir aleatoriamente ocho talleres de reparación, la organización de consumidores hizo que el hombre llevara su coche a cuatro de ellos y la mujer a los otros cuatro. Una semana más tarde se repitió el proceso, yendo el hombre a los talleres a los que previamente había ido la mujer y viceversa. Los presupuestos en dólares recibidos fueron los siguientes:

Taller	Presupuestos dados al hombre (en $)	Presupuestos dados a la mujer (en $)
1	145	145
2	220	300
3	150	200

(*Continúa*)

Taller	Presupuestos dados al hombre (en $)	Presupuestos dados a la mujer (en $)
4	100	125
5	250	400
6	150	135
7	180	200
8	240	275

Contraste la hipótesis de que el sexo de la persona que lleva el coche al taller de reparación no afecta al presupuesto entregado, cuando se utiliza

(a) Contraste de signos

(b) Contraste de rangos signados

8. Se trató a once pacientes con un alto contenido de albúmina en sangre con un medicamento específico. Los valores de albúmina medidos antes y después de la medicación fueron los siguientes:

Contenido de albúmina en sangre (en gramos por 100 ml)

Paciente	Antes de la medicación	Después de la medicación
1	5,04	4,82
2	5,16	5,20
3	4,75	4,30
4	5,25	5,06
5	4,80	5,38
6	5,10	4,89
7	6,05	5,22
8	5,27	4,69
9	4,77	4,52
10	4,86	4,72
11	6,14	6,26

(a) ¿Cuál es el estadístico del contraste de rangos signados?

(b) Si se contrasta si el tratamiento tiene o no efecto, ¿cuál es el p valor resultante?

9. Un ingeniero mantiene que pintar el exterior de un tipo determinado de avionetas afecta a su velocidad de crucero. Para contrastar esta idea se probaron 10 avionetas de ese tipo, que acababan de salir de la cadena de montaje, para determinar su velocidad de crucero antes y después de que se pintaran. Resultaron los datos siguientes:

	Velocidad de crucero (en millas por hora)	
Avioneta	**Antes de pintarla**	**Después de pintarla**
1	426,1	416,7
2	438,5	431,0
3	440,6	442,6
4	418,5	423,6
5	441,2	447,5
6	427,5	423,9
7	412,2	412,8
8	421,0	419,8
9	434,7	424,1
10	411,9	418,7

¿Prueban estos datos que la idea del ingeniero es correcta? Utilice un nivel de significación del 5%.

10. Sea $X_1, \ldots, X_n$ una muestra aleatoria extraída de una determinada población. Supongamos que se desea contrastar la hipótesis de que la distribución poblacional es simétrica respecto de un valor dado ν. Explique cómo esto se puede llevar a cabo con el contraste de rangos signados. (*Sugerencia*: Haga $D_i = X_i - \nu$.)

14.4 Contraste de la suma de rangos para comparar dos poblaciones

Consideremos dos poblaciones sobre las que incide una determinada característica cuantitativa y supongamos que se desea contrastar la hipótesis de que las dos distribuciones poblacionales de esa característica son idénticas. Para contrastar esta hipótesis se extraen dos muestras independientes de tamaños n y m, respectivamente, de la primera y la segunda población.

Si se asume que las distribuciones en las dos poblaciones son normales se podrían utilizar los contrastes con dos muestras desarrollados en el capítulo 10. Sin embargo, a continuación presentaremos un contraste no paramétrico que no requiere la asunción de la normalidad.

Para empezar ordene de menor a mayor los $n + m$ valores de las dos muestras. Esto es, asigne el rango 1 al menor valor, el rango 2 al segundo menor valor, y así sucesivamente. Por el momento, asumiremos que todos los $n + m$ valores son distintos y que, por tanto, no se producen empates. Designemos una de las muestras (no importa cuál) como primera muestra. El contraste que se propone utiliza como estadístico del contraste la suma de los rangos de la primera muestra. Esto es,

$$\text{TS} = \text{suma de rangos de los datos de la primera muestra}$$

Ejemplo 14.8 Para determinar si el tiempo de respuesta refleja a un estímulo depende de la edad se seleccionaron dos muestras independientes: una de varones de 20 años de edad y otra de varones de 50. Los siguientes datos recogen sus tiempos de reacción (en segundos) al estímulo.

Varones de 20 años de edad: 4,22, 5,13, 1,80, 3,34, 2,72, 2,80, 4,33, 3,60
Varones de 50 años de edad: 5,42, 3,39, 2,55, 4,45, 5,55, 4,96, 5,88, 6,30, 5,10

Si estos 17 datos se escriben en orden creciente resulta:

1,80,* 2,55, 2,72,* 2,80,* 3,34,* 3,39, 3,60,* 4,22,* 4,33,*
4,45, 4,96, 5,10, 5,13,* 5,42, 5,55, 5,88, 6,30

Se ha incluido un asterisco a continuación de los datos que proceden de la muestra de varones de 20 años (que será tomada como primera muestra). De aquí se desprende que la suma de rangos de la primera muestra es

$$\text{TS} = 1 + 3 + 4 + 5 + 7 + 8 + 9 + 13 = 50 \quad \blacksquare$$

Sea H_0 la hipótesis de que las dos distribuciones poblacionales son idénticas, y supongamos que el valor del estadístico del contraste es t. Puesto que se debe rechazar H_0 si el valor de TS es o significativamente grande o significativamente pequeño, se tiene que el contraste al nivel de significación a propondrá rechazar H_0 bien si

$$P\{\text{TS} \leq t\} \leq \frac{\alpha}{2}$$

o bien si

$$P\{\text{TS} \geq t\} \leq \frac{\alpha}{2}$$

donde las dos probabilidades se deben calcular asumiendo que H_0 es cierta. Dicho de otro modo, se rechazará la hipótesis nula si la suma de rangos de la primera muestra es bien demasiado pequeña o bien demasiado grande para que pueda deberse al azar. Como resultado se desprende que, a nivel de significación a, el contraste propondrá rechazar H_0 si el p valor resultante, dado por

$$p \text{ valor} = 2 \text{ Min}(P\{\text{TS} \leq t\}, P\{\text{TS} \geq t\})$$

es menor o igual que a.

Para calcular las probabilidades involucradas se necesita conocer la distribución de TS cuando H_0 es cierta. Para empezar, supongamos que la primera muestra es la que tiene un tamaño n. Si H_0 es cierta y, por tanto, los valores $n + m$ de datos proceden de la misma distribución se tiene que el conjunto de los rangos de la primera muestra tendrá la misma distribución que la de una selección aleatoria de n de los valores $1, 2, \ldots, n + m$. Con ello, se puede demostrar que, si H_0 es cierta, la media y la varianza de TS vienen dadas por las expresiones que se indican a continuación.

Si H_0 es cierta,

$$E[\text{TS}] = \frac{n(n + m + 1)}{2}$$

$$\text{Var}(\text{TS}) = \frac{nm(n + m + 1)}{12}$$

Adicionalmente, se puede demostrar que, cuando n y m están por encima de un umbral moderado (ambos por encima de 7 deberían ser suficientes), TS sigue aproximadamente una distribución normal, cuando H_0 es cierta. De aquí se desprende que, si los tamaños muestrales no son demasiado pequeños, TS se distribuye aproximadamente como una normal con la media y la varianza anteriormente indicadas.

Ejemplo 14.9 En el ejemplo 14.8 se verifica que $n = 8$, $m = 9$ y que el valor de la suma de los rangos de la primera muestra es TS = 50. Se tiene además que

$$E[\text{TS}] = \frac{n(n + m + 1)}{2} = 72 \qquad \text{Var}(\text{TS}) = \frac{nm(n + m + 1)}{12} = 108$$

Puesto que el valor observado de TS es menor que su media, se tiene que

$$\begin{aligned} p \text{ valor} &= 2P\{\text{TS} \leq 50\} \\ &= 2P\{\text{TS} \leq 50{,}5\} \\ &= 2P\left\{\frac{\text{TS} - 72}{\sqrt{108}} \leq \frac{50{,}5 - 72}{\sqrt{108}}\right\} \\ &\approx 2P\{Z \leq -2{,}069\} \\ &= 0{,}0385 \end{aligned}$$

En consecuencia, la hipótesis nula de que las dos distribuciones poblacionales son idénticas se debe rechazar al nivel de significación del 5%. ■

Si existieran empates, el rango de un valor muestral de empate se asignará de forma que coincida con la media de todos los datos con el mismo valor. Por ejemplo, si los datos de la primera muestra fueran 2, 4, 4 y 6, y los de la segunda 5, 6 y 7, la suma de rangos de los datos de la primera muestra sería $1 + 2{,}5 + 2{,}5 + 5{,}5 = 11{,}5$. Posteriormente, el contraste se realizará exactamente igual que antes. (Sucede que la posibilidad de empates tiene un efecto reductor de la varianza de TS cuando la hipótesis nula es cierta. Como resultado, el p valor indicado anteriormente será mayor que el p valor real que tiene en cuenta los empates. En consecuencia, el contraste presentado inicialmente es conservador, en el sentido de que siempre que se presenten empates y que se rechace H_0 con el contraste inicial, el contraste más sofisticado que promedia los rangos de los empates también rechazará la hipótesis nula.)

La Estadística en perspectiva

El contraste presentado en este apartado se denomina *contraste de la suma de rangos con dos muestras*. Dejando aparte el contraste de signos, que se remonta a 1710 y se debe a Arbuthnot (véase la sección 9.5), fue uno de los primeros contrastes no paramétricos que se desarrollaron. Fue propuesto, a mediados de década de 1940, independientemente por Wilcoxon y el equipo de Mann y Whitney. Por este motivo, en unas ocasiones, se denomina *contraste de la suma de rangos de Wilcoxon* y, en otras, *contraste de Mann-Whitney*. Las publicaciones de Mann-Whitney fueron el inicio de una oleada de investigaciones sobre los contrastes no paramétricos que aún continúa en la actualidad.

Ejemplo 14.10 Para determinar si las habilidades con el lenguaje de dos estudiantes eran similares, un profesor de Lengua propuso a sus alumnos que redactaran un trabajo sobre el mismo tema. Después, el profesor contó el número de palabras con más de cuatro letras que había empleado cada estudiante. Los resultados obtenidos fueron los siguientes:

	Número de palabras usadas con más de *i* letras	
i	**Estudiante 1**	**Estudiante 2**
4	44	49
5	16	11
6	8	5
7	7	4
8	4	1
9	2	1
10	3	0

Así, por ejemplo, 8 de las 84 palabras (con más de 4 letras) escritas por el estudiante 1 y 5 de las 71 usadas por el estudiante 2 tenían exactamente 6 letras. Utilice estos datos para contrastar la hipótesis de que las distribuciones de frecuencias de las palabras con más de tres letras son iguales para los dos estudiantes.

Solución Los datos consisten en dos muestras, una de 84 palabras y la otra de 71. En la muestra combinada de 155 palabras, el valor 4 aparece 93 veces; por consiguiente, a estos 93 datos se les otorgará un rango igual a la media de los números naturales del 1 al 93, es decir,

$$\frac{1 + 2 + \cdots + 93}{93} = \frac{1 + 93}{2} = 47$$

De igual forma, puesto que el siguiente menor valor de datos (el valor 5) ocurre en 27 ocasiones, cada uno de estos datos compartirán un rango igual a la media de los enteros comprendidos entre 94 y 120, esto es,

$$\frac{94 + 120}{2} = 107$$

Si se actúa igual con las palabras de 6, 7, 8, 9 y 10 letras, se obtienen los rangos que se asignarán a los restantes valores distintos de datos. Todo ellos se reflejan a continuación:

Valor de dato	4	5	6	7	8	9	10
Rango	47	107	127	139	147	151	154

La suma de rangos de la muestra de 71 palabras toma el valor

$$\text{TS} = 47 \times 49 + 107 \times 11 + 127 \times 5 + 139 \times 4 + 147 + 151 = 4969$$

Puesto que $n = 71$ y $m = 84$, se tiene que

$$\frac{n(n + m + 1)}{2} = 5538 \qquad \frac{nm(n + m + 1)}{12} = 77\,532$$

Así pues, el p valor aproximado es

$$\begin{aligned} p \text{ valor} &= 2P\{\text{TS} \leq 4969\} \\ &= 2P\{\text{TS} \leq 4969{,}5\} \\ &= 2P\left\{\frac{\text{TS} - 5538}{\sqrt{77\,532}} \leq \frac{4969{,}5 - 5538}{\sqrt{77\,532}}\right\} \\ &\approx 2P\{Z \leq -2{,}04\} \\ &= 0{,}041 \end{aligned}$$

y, por consiguiente, se debe rechazar la hipótesis de que las distribuciones de las longitudes de palabras usadas por los dos estudiantes son iguales, al nivel de significación del 5%. ■

Si los tamaños muestrales no son grandes —es decir, si alguno de ellos es menor que 8— no se puede asumir que la distribución de la suma de rangos sea aproximadamente normal. Sin embargo, aún así se puede utilizar el contraste de la suma de rangos si se calcula el p valor exacto. Para hacerlo se utiliza el hecho de que, cuando H_0 es cierta, el conjunto de rangos de la primera muestra sigue la misma distribución que se obtendría si se seleccionaran aleatoriamente n enteros del conjunto de valores $1, 2, \ldots, n + m$. Si tenemos esto en cuenta es posible —con la ayuda de un ordenador— determinar explícitamente el p valor. El Programa 14-2 permite obtener el p valor exacto para un contraste de la suma de rangos concreto. Las entradas que se suministrarán al programa son los tamaños muestrales de la primera y la segunda muestra y la suma de los rangos de los elementos de la primera muestra. Aunque cualquiera de las dos muestras puede ser considerada como primera muestra, el programa es más rápido si se considera que la primera muestra es aquella cuya suma de rangos es menor. Adicionalmente, dado que este programa asume implícitamente que no se producen empates, tan sólo se puede utilizar cuando el valor de TS es entero.

Perspectiva histórica

(Ohio State University Photo Archive, Columbus)

Thomas Mendenhall

El uso de técnicas estadísticas para llevar a cabo comparaciones literarias se inició hace muchos años. En 1901, Thomas Mendenhall, que había sido profesor de física en la Universidad del Estado de Ohio, publicó un estudio frecuentista que comparaba las longitudes de las palabras usadas por Shakespeare y por otros autores. Mendenhall observó que casi todas las obras de teatro de Shakespeare tenían aproximadamente la misma distribución. Comprobó, además, que Shakespeare, comparado con Dickens o Thackeray, usaba una mayor proporción de palabras cortas (con una, dos, tres, cuatro o cinco letras) y una menor proporción de palabras largas que ellos. Encontró también que la distribución de Francis Bacon era muy diferente de la de Shakespeare. Sin embargo, el entusiasmo inicial que produjo este hecho disminuyó cuando se comprobó que el análisis de las obras teatrales de Christopher Marlowe daba como resultado que su distribución de frecuencias de longitudes de palabras era casi idéntica a la de Shakespeare.

Más recientemente se empleó el análisis estadístico para decidir la autoría de los 12 *Artículos Federalistas*. Estos artículos, compuestos por 77 cartas, aparecieron de forma anónima en periódicos del Estado de Nueva York entre 1787 y 1788. Las cartas intentaban convencer a los ciudadanos de Nueva York para que ratificaran la Constitución. Aunque se sabía con generalidad que los autores de los artículos fueron Alexander Hamilton, John Jay y James Madison, no se conocía quién de ellos había sido el autor de cada uno de los artículos. En 1964, ya se había asignado la autoría de la mayor parte de los artículos. Sin embargo continuaba habiendo una encarnizada disputa acerca de la autoría de 12 de ellos. En un libro publicado en 1964, los estadísticos de Harvard Frederic Mosteller y David Wallance, mediante análisis estadísticos, concluyeron que el total de los 12 habían sido escritos únicamente por Madison. El citado análisis tuvo en cuenta cuestiones como las distribuciones de frecuencias del uso que cada autor hacía de palabras tales como *by, from, to* y *upon* (por, desde, a y sobre).

Ejemplo 14.11 Consideremos el ejemplo 14.9, utilizando en esta ocasión el Programa 14-2 para computar el p valor. Este programa se ejecuta más rápidamente si se identifica como primera muestra aquella que tiene menor suma de rangos. El tamaño de la primera muestra es 8, y el de la segunda 9. La suma de los rangos de la primera muestra es 50. El Programa 14-2 calcula como p valor 3,595229E-02.

Así pues, el p valor exacto es 0,0359, que es razonablemente próximo al valor 0,0385 obtenido con la aproximación normal. ■

14.4.1 Comparación de los contrastes no paramétricos con los contrastes que asumen la distribución normal

El punto fuerte de los contrastes no paramétricos se basa en que se pueden utilizar sin asumir forma alguna acerca de las distribuciones subyacentes. El precio que se debe pagar cuando se usan contrastes no paramétricos es que, cuando las distribuciones subyacentes son normales o aproximadamente normales, no resultan ser tan efectivos como los contrastes vistos que presuponen la normalidad. Sin embargo, lo que resulta sorprendente es que la pérdida de efectividad es relativamente pequeña. Por ejemplo, se puede demostrar que, cuando los tamaños muestrales n y m son grandes, la eficiencia del contraste no paramétrico de la suma de

rangos es aproximadamente del orden del 95% de la eficiencia del contraste de la t con dos muestras, cuando las distribuciones poblacionales son verdaderamente normales. Grosso modo, esto significa que, cuando la distribuciones poblacionales son normales pero distintas, la probabilidad de rechazo de los contrastes no paramétricos con muestras de tamaño n es igual a la de los contrastes de la t, basados en la normalidad, de tamaño 0,95 n. Este resultado es sorprendente y nos lleva a concluir que los contrastes no paramétricos son mejores si uno no está absolutamente seguro de que las distribuciones subyacentes son próximas a la normal. Esto se debe a que, si las distribuciones no son normales, el contraste que asume la normalidad se basa en una hipótesis falsa; y, además, si las distribuciones son normales, los contrastes no paramétricos son casi tan buenos como los basados en la normalidad. Pese a lo anterior, también es cierto que, cuando las distribuciones subyacentes no son normales, los contrastes basados en la normalidad se comportan bien cuando los tamaños muestrales n y m son grandes. Esto ocurre porque los contrastes basados en la normalidad se basan en estadísticos del contraste cuya distribución es aproximadamente normal, incluso aunque las distribuciones poblacionales no lo sean. Se puede, pues, concluir que, para tamaños de muestras grandes, los contrastes de la t, pese a que presuponen la normalidad, son efectivos.

Probablemente lo mejor que se puede decir es que, si uno no está seguro de que las distribuciones subyacentes sean normales, los contrastes no paramétricos de la suma de rangos son preferibles a los contrastes de la t con dos muestras, si las muestras son de tamaño moderado. Por otro lado, cuando se dispone de muestras grandes, se pueden utilizar los dos tipos de contrastes. Sin embargo, una diferencia clave entre ellos, que es útil a la hora de decidir qué tipo de contraste se debe usar, es que los contrastes de la t están diseñados para detectar diferencias entre las medias poblacionales, mientras que los contrastes de la suma de rangos están diseñados para detectar diferencias entre las distribuciones poblacionales.

Problemas

1. Los siguientes datos provienen de muestras independientes extraídas de dos poblaciones.

   Muestra 1: 142, 155, 237, 244, 202, 111, 326, 334, 350, 247

   Muestra 2: 212, 277, 175, 138, 341, 255, 303, 188

   (a) Determine la suma de los rangos de los datos de la muestra 1.

   (b) Determine la suma de los rangos de los datos de la muestra 2.

2. Existe una identidad algebraica que establece que la suma de los k primeros enteros positivos es igual a $k(k + 1)/2$. Esto es,

$$\sum_{i=1}^{k} i = \frac{k(k + 1)}{2}$$

   Utilice esta identidad para obtener la relación entre la suma de los rangos de la muestra de tamaño n y la suma de los rangos de la muestra de tamaño m, asuma que todos los valores de datos $n + m$ son distintos. Compruebe los resultados del problema 1 a partir de lo anterior.

3. Se ha llevado a cabo un estudio para determinar si los rendimientos educacionales son iguales en las áreas rurales y en las áreas urbanas de California. Para ello, se han seleccionado dos áreas, una urbana y otra rural, con aproximadamente las mismas condicio-

nes socioeconómicas y, dentro de cada una de ellas, se han extraído dos muestras de estudiantes que estaban a punto de terminar su educación secundaria. Todos estos estudiantes realizaron un test de aptitud, cuyos resultados se muestran a continuación:

Área rural	Área urbana
544	610
567	498
475	505
658	711
590	545
602	613
571	509
502	514
578	609

Encuentre el p valor al contrastar la hipótesis de que las distribuciones de las calificaciones obtenidas en ambas áreas son idénticas. Utilice la aproximación normal.

4. Se dividió aleatoriamente a un grupo de 16 voluntarios en dos subgrupos de 8 personas. A los miembros del primer subgrupo se les suministró una dosis diaria de 5 gramos de vitamina C, mientras que a los del segundo se les suministró un placebo. Tras un mes de prueba, se midieron los niveles de colesterol en la sangre de los 16 voluntarios para compararlos con los niveles que tenían antes de que se realizara el experimento. Las reducciones de los niveles de colesterol para los dos subgrupos fueron las siguientes.

Vitamina C	Placebo
6	9
12	−3
14	0
2	−1
7	5
7	3
1	−4
8	−1

Al nivel de significación del 5%, contraste la hipótesis nula de que la vitamina C y el placebo son igualmente efectivos como reductores del colesterol. Asuma que, si la hipótesis nula es cierta, la distribución del estadístico del contraste es aproximadamente normal. (Un valor de dato negativo significa que el nivel de colesterol aumentó. Por ejemplo, el valor −4 indica que la medida tras la prueba resultó ser 4 unidades superior a la registrada antes de la prueba.)

5. Se ha llevado a cabo un estudio para contrastar la hipótesis de que las distribuciones de los salarios iniciales de los graduados en Informática por la Universidad de Stanford y por la Universidad de Berkeley son iguales. Los salarios anuales (en unidades de

1000 $) de los miembros de dos muestras de estudiantes recientemente graduados fueron los siguientes:

Stanford	Berkeley
57,8	52,6
60,4	56,6
71,2	61,0
52,5	47,9
68,0	55,0
69,6	62,5
70,0	66,4
54,0	57,5
48,8	56,5
57,6	49,8

Al nivel de significación del 5%, ¿qué conclusión se puede extraer?

6. Se ha llevado a cabo un experimento para determinar la efectividad de la vitamina B1 para estimular el crecimiento de los champiñones. Se aplicó la vitamina a 9 champiñones seleccionados aleatoriamente de un grupo de 17. A los restantes 8 champiñones se los mantuvo sin tratamiento. Los pesos (en gramos) de los 17 champiñones del experimento fueron los siguientes:

Champiñones no tratados: 18, 12,4, 13,5, 14,6, 24, 21, 23, 17,5
Champiñones tratados con vitamina B1: 34, 27, 21,2, 29, 20,5, 19,6, 28, 33, 19

Contraste la hipótesis, al nivel de significación del 5%, de que el tratamiento con vitamina B1 no es efectivo.

7. Se ha dividido aleatoriamente a un grupo de 24 trabajadores en dos conjuntos de 12 cada uno. Los trabajadores de cada conjunto fueron sometidos a un mismo programa de entrenamiento de 2 semanas de duración. Sin embargo, los miembros del primer conjunto tuvieron una reunión previa de un día para fomentar su motivación. Al final del periodo de entrenamiento, todos los trabajadores realizaron una serie de pruebas, y se ordenaron de acuerdo con las calificaciones globales que obtuvo cada uno. Si la suma de los rangos de los trabajadores que asistieron a la sesión de motivación fue 136, ¿cuál es el p valor obtenido al contrastar la hipótesis de que la sesión de motivación no tuvo efecto alguno?

8. Utilice el Programa 14-2 para obtener el p valor exacto en el problema 4.

9. Utilice el Programa 14-2 para calcular el p valor exacto en el problema 5.

10. Rehaga el problema 1 de la sección 10.4, y en esta ocasión utilice un contraste no paramétrico.

14.5 Contraste de rachas para la aleatoriedad

En gran parte de los análisis de datos se asume de antemano que el conjunto de datos se ha obtenido a partir de una muestra aleatoria procedente de una población determinada. Sin

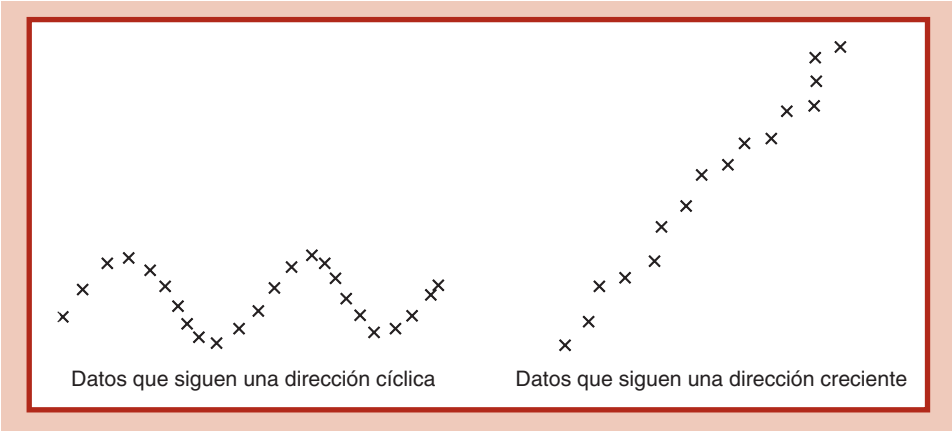


Figura 14.4 Conjuntos de datos no aleatorios.

embargo, en ocasiones ocurre que el conjunto de datos no constituye una muestra aleatoria procedente de una población, sino que por el contrario presenta un determinado patrón de comportamiento interno. Por ejemplo, los datos podrían tender a crecer o decrecer a lo largo del tiempo, o podrían seguir cierto patrón cíclico según el cual primero crecen y luego decrecen de forma periódica (véase la figura 14.4). En este apartado se desarrollará un contraste de la hipótesis de que un conjunto de datos dado constituye una muestra aleatoria.

Para contrastar la hipótesis de que una sucesión dada de valores de datos constituye una muestra aleatoria, se supondrá inicialmente que cada dato puede tomar únicamente dos valores posibles, que designaremos por 0 y 1. Consideremos un conjunto cualquiera de ceros y unos, y llamemos racha a cualquier sucesión de varios ceros consecutivos o de varios unos seguidos. Por ejemplo, el conjunto de datos

$$0, 1, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1, 1, 0, 1, 0, 0$$

contiene un total de 9 rachas: 5 rachas de ceros y 4 rachas de unos. La primera racha consiste en un único valor 0; la segunda racha la forma los tres valores 1, 1, 1 siguientes; la racha posterior está formada por los dos valores 0, 0; y así sucesivamente.

Supongamos que el conjunto de datos está formado por un total de $n + m$ valores, de los que n son iguales a 1 y m son iguales a 0. Denotemos por R el número de rachas del conjunto de datos. Si el conjunto de datos fuera una muestra procedente de una población, todas las posibles ordenaciones de los valores $n + m$ (con n unos y m ceros) serían igualmente probables. Con este resultado, es posible determinar la distribución de probabilidad de R y, de esta forma, llegar a contrastar la hipótesis nula H_0 de que el conjunto de datos es una muestra aleatoria, y se rechazará H_0 si el valor de R es demasiado grande o demasiado pequeño como para que se deba meramente al azar. Concretamente, si el valor de R es r, el contraste, a nivel de significación α, rechazará H_0 bien si

$$P\{R \leq r\} \leq \frac{\alpha}{2}$$

o bien si

$$P\{R \geq r\} \leq \frac{\alpha}{2}$$

donde estas probabilidades se deben calcular asumiendo que la hipótesis nula es cierta. El contraste resultante se denomina *contraste de rachas*.

También se puede llevar a cabo el contraste de rachas observando el valor de R, es decir, conociendo r y calculando, después, el p valor resultante:

$$p \text{ valor} = 2 \text{ Min}(P\{R \leq r\}, P\{R \geq r\})$$

El Programa 14-3 obtiene este p valor a partir del cálculo de las probabilidades involucradas.

Ejemplo 14.12 La siguiente sucesión muestra los resultados de los últimos 24 partidos jugados por el equipo local de una ciudad. Las letras W y L significan que el equipo local ha ganado y ha perdido, respectivamente.

$$W, L, L, L, W, L, L, W, L, L, W, L, L, W, L, W, L, L, L, L, W, L, W, L$$

¿Son estos datos consistentes con la aleatoriedad?

Solución Para contrastar la hipótesis de aleatoriedad, observe que el conjunto de datos de 8 W y 16 L contiene 16 rachas. Para ver si esto justifica que se rechace la hipótesis de aleatoriedad, utilizaremos el Programa 14-3.

El número de unos es 8 y el de ceros 16. El número de rachas es 16. El p valor proporcionado por el Programa 14-3 es 0,052497.

Puesto que el p valor es 0,0525, se desprende que no se puede rechazar la hipótesis de aleatoriedad, al nivel de significación del 5%. Es decir, aunque la evidencia está en contra de la hipótesis de aleatoriedad, ésta no es lo suficientemente fuerte como para que se deba rechazar dicha hipótesis, al nivel de significación del 5%. ■

Si el Programa 14-3 no está disponible, se puede obtener un p valor aproximado si se usa un resultado que establezca que, cuando la hipótesis nula es cierta, R sigue aproximadamente una distribución normal con una media y unas varianza dadas, respectivamente, por

$$\mu = \frac{2nm}{n + m} + 1$$

y

$$\sigma^2 = \frac{2nm(2nm - n - m)}{(n + m)^2 (n + m - 1)}$$

Esta aproximación de la distribución de R es aceptable siempre que n y m, los números de unos y ceros del conjunto de datos, sobrepasen ambos un determinado umbral mínimo. (Es suficiente que ambos sean mayores que 20.)

Si se han observado r rachas en total, el p valor viene dado por

$$p \text{ valor} = 2 \text{ Min}(P\{R \leq r\}, P\{R \geq r\})$$

Se puede utilizar dicha aproximación normal para calcular las probabilidades incluidas en la expresión anterior.

Ejemplo 14.13 Repitamos el ejemplo 14.12, utilizando en esta ocasión la aproximación normal para obtener el p valor. Puesto que $n = 8$ y $m = 16$, se ve que

$$\mu = \frac{2 \cdot 8 \cdot 16}{24} + 1 = 11{,}667$$

$$\sigma^2 = \frac{256(256 - 24)}{24 \cdot 24 \cdot 23} = 4{,}4831$$

Existe un total de 16 rachas, de donde el p valor viene dado por

$$p \text{ valor} = 2 \text{ Min}(P\{R \leq 16\}, P\{R \geq 16\})$$

Como $E[R] = 11{,}667$ es menor que 16, se tiene que $P\{R \geq 16\}$ es menor que $P\{R \leq 16\}$. Por consiguiente,

$$\begin{aligned} p \text{ valor} &= 2P\{R \geq 16\} \\ &= 2P\{R \geq 15{,}5\} \quad \text{(corrección por continuidad)} \\ &= 2P\left\{\frac{R - 11{,}667}{\sqrt{4{,}4831}} \geq \frac{15{,}5 - 11{,}667}{\sqrt{4{,}4831}}\right\} \\ &\approx 2P\{Z \geq 1{,}81\} \\ &= 0{,}07 \end{aligned}$$

Puesto que este p valor aproximado es mayor que 0,05, no se puede rechazar la hipótesis nula, al nivel de significación del 5%. Así pues, el p valor aproximado nos lleva a la misma conclusión, al nivel de significación del 5%, que el p valor exacto. Pese a ello, el p valor aproximado de 0,07 no está muy próximo al real de 0,0525. Es evidente que, en este ejemplo, los valores de n y m (iguales a 8 y 16) no verifican la regla empírica indicada, en el sentido de que ambos deben ser como mínimo 20 para que la aproximación sea precisa. ■

Ejemplo 14.14 Consideremos una sucesión de 20 ceros y 20 unos en la que existen 27 rachas. Compare el p valor real con el aproximado, al contrastar de la hipótesis de aleatoriedad de la sucesión.

Solución Se empezará utilizando la aproximación normal. Puesto que $n = m = 20$, la media y la desviación típica del número de rachas son, respectivamente,

$$\mu = \frac{2 \cdot 20 \cdot 20}{40} + 1 = 21 \qquad \sigma = \sqrt{\frac{2 \cdot 20 \cdot 20 \cdot 760}{40 \cdot 40 \cdot 39}} = 3{,}121$$

Por consiguiente, dado que el número observado de rachas es 27, la aproximación normal conduce a:

$$\begin{aligned} p \text{ valor} &= P\{R \geq 27\} \\ &= 2P\{R \geq 26{,}5\} \\ &= 2P\left\{\frac{R - 21}{3{,}121} \geq \frac{26{,}5 - 21}{3{,}121}\right\} \\ &\approx 2P\{Z \geq 1{,}762\} \\ &= 0{,}078 \end{aligned}$$

Por otro lado, con el Programa 14-3 se obtiene el p valor exacto:

El número de unos es 20. El número de ceros es 20. El número de rachas es 27. El p valor computado con el Programa 14-3 es 0,075996.

Por consiguiente, en este ejemplo (donde tanto n como m son iguales a 20) el p valor aproximado, 0,078, es bastante próximo al p valor real, 0,076. ■

También se puede utilizar el contraste de rachas para contrastar la aleatoriedad cuando la sucesión de valores no se compone únicamente de ceros y unos. Para contrastar si una sucesión dada de datos $X_1, X_2, \ldots, X_n$ constituye una muestra aleatoria procedente de una determinada población, denotemos por s_m la mediana muestral del conjunto de datos. Para cada valor, determinemos si es menor o igual que s_m o si es mayor que s_m y escribamos en la posición i un 0, si X_i es menor o igual que s_m, o un 1, en caso contrario. Si el conjunto de datos original constituye una muestra aleatoria procedente de una distribución, también las sucesiones de 0 y 1 constituirán una muestra aleatoria. Por consiguiente, se puede contrastar si el conjunto de datos original es una muestra aleatoria si se utiliza el contraste de rachas sobre la sucesión resultante de ceros y unos.

Ejemplo 14.15 Las temperaturas medias en verano, medidas en grados Fahrenheit, de los 20 años comprendidos entre 1971 y 1990 en una ciudad de la costa oeste de Estados Unidos fueron

72, 71, 70, 82, 80, 77, 71, 85, 75, 80, 82, 81, 83, 82, 85, 86, 83, 81, 82, 84

Contraste la hipótesis de que los datos constituyen una muestra aleatoria.

Solución La mediana muestral es el promedio de los valores que ocupan las posiciones 10 y 11 una vez que los datos se han ordenado de menor a mayor. Por consiguiente,

$$s_m = \frac{81 + 82}{2} = 81{,}5$$

La sucesiones de ceros y unos, que indican si un determinado valor es menor o igual que 81,5 o mayor que este valor, coinciden con

0 0 0 1 0 0 0 1 0 0 1 0 1 1 1 1 1 0 1 1

La sucesión se compone de 10 ceros y 10 unos, y consta de 10 rachas. Para determinar si este valor es significativamente muy grande o muy pequeño como para que pueda deberse al azar si los datos fueran realmente una muestra aleatoria, se usará el Programa 14-3.

El número de unos y de ceros es 10, en ambos casos. El número de rachas es igualmente 10. El p valor obtenido es 0,8281409.

Para un p valor tan grande como éste, la hipótesis de aleatoriedad no se puede rechazar. Esto es, los datos no proporcionan ninguna evidencia de que no sean una muestra aleatoria.

Si se hubiera utilizado la aproximación normal, deberíamos haber empezado computando μ y σ, la media y la desviación típica del número total de rachas, bajo la hipótesis nula. Puesto que $n = m = 10$,

$$\mu = \frac{200}{20} + 1 = 11 \qquad \sigma = \sqrt{\frac{200(180)}{400 \cdot 19}} = 2{,}176$$

Finalmente, dado que el número observado de rachas es 10, el p valor que se obtiene con la aproximación normal es

$$\begin{aligned} p \text{ valor} &= 2P\{R \leq 10\} \\ &= 2P\{R \leq 10{,}5\} \\ &= 2P\left\{\frac{R - 11}{2{,}176} \leq \frac{10{,}5 - 11}{2{,}176}\right\} \\ &\approx 2P\{Z \leq -0{,}23\} \\ &= 0{,}818 \end{aligned}$$

Se ve, pues, que la aproximación normal es bastante precisa en este caso. ■

Ejemplo 14.16 La sucesión siguiente refleja las puntuaciones obtenidas por un equipo de baloncesto de un instituto en 23 partidos jugados durante el año académico 1994/95. ¿Es razonable asumir que las puntuaciones constituyen una muestra aleatoria?

77, 62, 58, 64, 66, 72, 59, 69, 80, 74, 72, 69, 74, 83, 85, 87, 80, 88, 76, 77, 82, 85, 83

Solución La mediana muestral coincide con la 12ª menor puntuación, es igual a 76. La sucesión de ceros y unos, que indica si cada uno de los valores es menor o igual que 76 o mayor que 76, es las siguientes:

1 0 0 0 0 0 0 0 1 0 0 0 0 1 1 1 1 1 0 1 1 1 1

En conclusión, esta sucesión se compone de doce ceros y de once unos, y tiene 7 rachas. A partir del Programa 14-3, se obtiene que el p valor es de 0,02997; así pues, se debe rechazar la hipótesis de que los datos constituyen una muestra aleatoria, al nivel de significación del 5%. ■

Problemas

A menos que se diga lo contrario, utilice el Programa 14-3 o la aproximación normal, según sea más conveniente, para resolver los problemas siguientes.

1. Consideremos una sucesión de ceros y unos compuesta por veinte ceros y treinta unos. Sea R el número total de rachas. ¿Cuál es (a) el mayor y (b) el menor de los valores posibles de R?

2. Determine el número de rachas de los siguientes conjuntos de datos de ceros y unos.

   (a) 1 1 0 0 0 0 1 1 0 1 0 1 1 0 0 0 1 1

   (b) 0 1 1 0 0 0 0 1 1 1 0 1 1 0 1 1 1 1

   (c) 0 0 0 1 1 0 0 1 1 0 1 1 1 1 1 0 1 0

3. Los datos siguientes se refieren al nivel de calidad de los 26 últimos relojes producidos en una fábrica suiza. El valor 1 significa que la calidad es aceptable, y el valor 0 que es inaceptable.

   1 1 1 0 1 1 1 1 1 1 1 1 1 0 0 0 0 1 1 1 1 0 0 1 1

   Contraste la hipótesis, al nivel de significación del 5%, de que los datos constituyen una muestra aleatoria.

4. En una cadena de producción se han fabricado 60 artículos, de los cuales 12 resultaron ser defectuosos. Los artículos defectuosos fueron los fabricados en las posiciones 9, 14, 15, 26, 30, 36, 37, 44, 45, 46, 59 y 60.

   (a) ¿Cuál es el valor de R y el número total de rachas?

   (b) Al nivel de significación del 5%, ¿se puede concluir que los sucesivos artículos fabricados constituyen una muestra aleatoria?

5. Se va a entrevistar a un total de 25 personas, 10 de las cuales son mujeres. Al entrevistador se le indica que las entreviste en un orden elegido aleatoriamente. Supongamos que la sucesión de los sexos de las personas entrevistadas es la siguiente (M = hombre; F = mujer):

   F F M F F F F M M F F M F F M M M M M M M M M M M

   ¿Siguió el entrevistador las instrucciones dadas? Explique por qué y calcule el p valor resultante.

6. A lo largo de los últimos 50 días, el índice Dow Johnes de producción industrial subió en 32 días y bajó en los 18 días restantes. Si el número total de rachas (de incremento o de disminución del índice Dow Johnes) fue de 22, ¿cuál es el p valor al contrastar la hipótesis de que los incrementos y las disminuciones constituyen una muestra aleatoria?

7. Las duraciones, en horas, de 30 baterías producidas sucesivamente fueron las siguientes:

   148, 152, 155, 147, 176, 170, 165, 149, 138, 155, 160, 153, 162, 155, 159,
   174, 168, 149, 182, 177, 191, 185, 178, 176, 182, 184, 181, 177, 160, 154

(a) ¿Cuál es la mediana muestral?

(b) ¿Cuál es el valor de R, número de rachas tras haber codificado cada dato dependiendo de si es menor (o igual) que o mayor que la mediana muestral?

(c) ¿Están estos datos en contra de la hipótesis de que la sucesión de valores constituye una muestra aleatoria?

8. Los siguientes datos representan los valores de cierre anual del índice Dow Johnes en 10 años consecutivos:

910, 890, 1010, 1033, 1080, 1275, 1288, 1553, 1980, 2702

Contraste la hipótesis, al nivel de significación del 5%, de que estos datos pueden constituir una muestra aleatoria.

Términos clave

Contrastes de hipótesis no paramétricos: Clase de contrastes de hipótesis sobre una población acerca de cuya distribución no se asume ningún tipo específico.

Contraste de signos: Contraste no paramétrico referido a la mediana de una población. El estadístico del contraste es igual al número de valores de datos que son menores que la mediana que se va a contrastar.

Contraste de rangos signados: Contraste no paramétrico cuya hipótesis nula es que la distribución poblacional es simétrica respecto de un determinado valor.

Contraste de la suma de rangos: Contraste no paramétrico acerca de la igualdad de dos distribuciones poblacionales. Parte de dos muestras independientes extraídas de cada una de las poblaciones, y luego utiliza los rangos que ocupan, en los datos combinados, los elementos de cada una de las dos muestras. El estadístico del contraste se define como la suma de los rangos de una (cualquiera) de las muestras.

Contraste de rachas: Contraste no paramétrico de la hipótesis de que una sucesión de datos constituye una muestra aleatoria procedente de una población.

Resumen

En este capítulo se introduce cómo se puede contrastar una hipótesis estadística sin asumir forma alguna sobre las distribuciones de probabilidad subyacentes. Dichos contrastes se dice que son *no paramétricos*.

Contraste de signos. El contraste de signos se puede utilizar para contrastar hipótesis relativas a la mediana de una distribución. Supongamos que, para un valor prefijado m, se desea contrastar

$$H_0: \eta = m$$

frente a

$$H_1: \eta \neq m$$

siendo η la mediana de la distribución poblacional. Para llevar a cabo este contraste se elige una muestra de elementos de la población y se eliminan todos los valores de datos que resulten ser iguales a m. Supongamos que el número de datos muestrales no eliminados es n. El estadístico del contraste de signos es el número de valores no eliminados que son menores que m. Si el número de tales valores es i, el p valor del contraste de signos viene dado por

$$p \text{ valor } = \begin{cases} 2P\{N \leq i\} & \text{si } i \leq \dfrac{n}{2} \\ 2P\{N \geq i\} & \text{si } i \geq \dfrac{n}{2} \end{cases}$$

donde N representa una variable aleatoria binomial de parámetros n y $p = 1/2$. El cálculo de las probabilidades de la binomial se puede llevar a cabo con del Programa 5-1 o con la aproximación de la binomial a la normal.

El contraste de signos también se puede utilizar para contrastar la hipótesis nula unilateral

$$H_0: \eta \leq m \qquad \text{frente a} \qquad H_1: \eta > m$$

Para ello se utiliza el mismo estadístico del contraste anterior; es decir, el número de datos menores que m. Si el valor del estadístico del contraste es i, el p valor viene dado por

$$p \text{ valor } = P\{N \leq i\}$$

donde, de nuevo, N representa una binomial de parámetros n y $p = 1/2$.

Si se desean contrastar las hipótesis unilaterales

$$H_0: \eta \geq m \qquad \text{frente a} \qquad H_1: \eta < m$$

el p valor, cuando existen i datos menores que m, es

$$p \text{ valor } = P\{N \geq i\}$$

siendo N una binomial de parámetros n y $p = 1/2$.

Como en todos los contrastes de hipótesis, se debe rechazar la hipótesis nula a cualquier nivel de significación mayor o igual que el p valor.

Contraste de rangos signados. El contraste de rangos signados se utiliza para contrastar la hipótesis de que una distribución es simétrica respecto de 0. En muchas aplicaciones, la población consiste en las diferencias entre los datos apareados. El contraste de rangos signados comienza con la selección de una muestra aleatoria de la población y elimina de ella los datos iguales a 0. Después ordena los restantes valores, cuyo número será identificado por n, en orden creciente de sus valores absolutos. El estadístico del contraste

es igual a la suma de los rangos de los valores de datos negativos. Si el valor del estadístico del contraste TS es t, el p valor viene dado por

$$p \text{ valor} = 2 \text{ Min}(P\{\text{TS} \leq t\}, P\{\text{TS} \geq t\})$$

donde las probabilidades se deben computar asumiendo que la hipótesis nula es cierta. El p valor se puede obtener con el Programa 14-1 o con hecho de que, cuando la hipótesis nula es cierta y n es suficientemente grande, TS sigue aproximadamente una distribución normal con media y varianza dadas, respectivamente, por

$$E[\text{TS}] = \frac{n(n+1)}{4} \qquad \text{Var(TS)} = \frac{n(n+1)(2n+1)}{24}$$

Contraste de la suma de rangos. El contraste de la suma de rangos se puede utilizar para contrastar la hipótesis nula de que dos distribuciones poblacionales coinciden, cuando se dispone de dos muestras independientes procedentes de cada una de las poblaciones. Designemos arbitrariamente a una cualquiera de las muestras como la primera. Supongamos que el tamaño de esa primera muestra es n y el de la otra es m. Ordenemos los miembros de las dos muestras combinadas. El estadístico del contraste de la suma de rangos coincide con la suma de los rangos de la primera muestra en la ordenación anterior. El contraste de la suma de rangos rechazará la hipótesis nula cuando el valor del estadístico del contraste sea significativamente grande o significativamente pequeño.

Cuando n y m son ambos mayores que 7 y se asume que H_0 es cierta, el estadístico del contraste sigue aproximadamente una distribución normal cuya media y varianza vienen, respectivamente, dadas por

$$E[\text{TS}] = \frac{n(n+m+1)}{2} \qquad \text{Var(TS)} = \frac{nm(n+m+1)}{12}$$

Cuando TS $= t$, lo anterior nos permite aproximar el p valor, que vendrá dado por

$$p \text{ valor} \approx \begin{cases} 2P\left\{Z \leq \dfrac{t + 0{,}5 - n(n+m+1)/2}{\sqrt{nm(n+m+1)/12}}\right\} & \text{si } t < \dfrac{n(n+m+1)}{2} \\ 2P\left\{Z \geq \dfrac{t - 0{,}5 - n(n+m+1)/2}{\sqrt{nm(n+m+1)/12}}\right\} & \text{si } t > \dfrac{n(n+m+1)}{2} \end{cases}$$

Para valores de t próximos a $n(n+m+1)/2$, el p valor es casi 1 y, en consecuencia, no se debería rechazar la hipótesis nula en ningún caso (no siendo necesario calcular las probabilidades anteriores).

Para valores pequeños de n y m se puede obtener el valor p exacto con el Programa 14-2.

Contraste de rachas. Se puede utilizar el contraste de rachas para contrastar la hipótesis nula de que una determinada sucesión de datos constituye una muestra aleatoria procedente de alguna población. Asume que cada dato es un 0 o un 1. En la sucesión de datos,

cualquier grupo de ceros o de unos consecutivos se denomina *racha*. El estadístico del contraste de rachas coincide con R, el número total de rachas. Si el valor observado de R es r, el p valor del contraste de rachas viene dado por

$$p \text{ valor} = 2 \text{ Min}(P\{R \leq r\}, P\{R \geq r\})$$

donde las probabilidades se deben calcular asumiendo que la hipótesis nula es cierta.

Se puede utilizar el Programa 14-3 para calcular la anterior expresión del p valor. Ahora bien, si este programa no está disponible, también se puede aproximar el p valor mediante el hecho de que, si la hipótesis nula es cierta, R sigue aproximadamente una distribución normal, cuya media y varianza son, respectivamente,

$$\mu = \frac{2nm}{n + m} + 1 \qquad \sigma^2 = \frac{2nm(2nm - n - m)}{(n + m)^2(n + m - 1)}$$

Problemas de repaso

1. Utilice un contraste no paramétrico para resolver el problema 2 de la sección 10.4.
2. Utilice un contraste no paramétrico para resolver el problema 3 de la sección 10.4.
3. De acuerdo con el *Boletín de la Reserva Federal* de Estados Unidos, de enero de 1992, la mediana muestral de los salarios netos de los trabajadores en 55 años fue, en 1989, igual a 104 500 $. (La media muestral fue, por su parte, igual a 438 300 $.) Supongamos que, en el año actual, se ha extraído una muestra de 1000 de tales trabajadores, de los cuales 421 tenían un salario neto superior a 104 500 $ (en dólares de 1989). Al nivel de significación del 5%, ¿se puede concluir que el salario neto mediano ha decrecido para dicho grupo de trabajadores?
4. Se ha diseñado un experimento para estudiar el efecto que tiene un nuevo detergente de gasolina sobre el consumo de los automóviles. Los datos siguientes representan el número de millas recorridas por galón antes y después de haber añadido el detergente en cuestión al combustible de ocho coches:

Coche	Millas sin aditivo	Millas con aditivo
1	24,2	23,5
2	30,4	29,6
3	32,7	32,3
4	19,8	17,6
5	25,0	25,3
6	24,9	25,4
7	22,2	20,6
8	21,5	20,7

Encuentre el p valor cuando se utilizan estos datos para contrastar la hipótesis nula de que el aditivo no afecta la distancia recorrida por galón, utilizando:

(a) el contraste de signos

(b) el contraste de rangos signados

Compare los resultados obtenidos, y éstos con los del ejemplo 10.8 de la sección 10.5.

5. Contraste la hipótesis de que los pesos de los estudiantes dados en el Apéndice A constituyen una muestra aleatoria. Utilice los primeros 40 valores de datos para contrastar esta hipótesis. ¿Cuál es el p valor resultante?

6. Seleccione una muestra aleatoria de 30 estudiantes del Apéndice A. Utilice sus datos para contrastar la hipótesis de que el peso mediano de todos los estudiantes de la lista es inferior o igual a 130 libras.

7. Seleccione una muestra aleatoria de 40 estudiantes del Apéndice A, y utilice esta muestra para contrastar la hipótesis de que la distribución de los niveles de colesterol es igual para ambos sexos.

8. Un químico lleva a cabo análisis de sangre buscando un determinado virus. Los sucesivos resultados obtenidos se muestran a continuación, donde p y A significan que el virus estaba presente y ausente, respectivamente.

$$A\ A\ p\ p\ p\ A\ A\ A\ A\ p\ p\ A\ A\ A\ A\ A\ A\ p\ p\ A\ A\ p\ p\ p\ P$$

Contraste la hipótesis de que el químico llevó a cabo los análisis siguiendo un orden aleatorio. Utilice un nivel de significación del 5%. Calcule, también, el p valor resultante.

9. Repita el ejemplo 10.9, y utilice en esta ocasión un contraste no paramétrico.

10. Explique cómo se puede analizar una tabla de contingencia para contrastar la hipótesis del ejemplo 14.10. Realice este contraste, calcule el p valor resultante y compare éste con el obtenido en el ejemplo 14.10. Dado que el contraste usual a partir de la tabla de contingencia es diferente del utilizado en el ejemplo 14.10, los dos p valores no tienen por qué coincidir.

11. Considere el problema 7 de la sección 14.3. Suponga ahora que se ha llevado el mismo coche a los 16 talleres de reparación, aunque en 8 de ellos fue la mujer quién solicitó el presupuesto y en los otros 8 fue el hombre. Suponga que las cantidades presupuestadas fueron las indicadas en el problema citado. Contraste la hipótesis, al nivel de significación del 5%, de que las distribuciones de los presupuestos entregados a los hombres y a las mujeres son iguales.

15

Control de calidad

El razonamiento estadístico será un día tan necesario para el ciudadano como la habilidad de leer y escribir.

H. G. Wells (1866–1946)

En este capítulo se presentarán los gráficos de control, que se utilizan para detectar cuándo un proceso de producción está fuera de control. Esto se puede hacer tanto cuando la variable de medida de los artículos producidos es continua como cuando es discreta. En el último caso, cada artículo producido puede tomar dos valores posibles, que identifican si el artículo es aceptable o no.

15.1 Introducción

En casi todos los sistemas industriales –incluyendo tanto los de producción o los de servicios– existe una variación aleatoria en los ítems procesados. Es decir, independientemente del rigor con que se esté controlando el proceso siempre habrá cierta variación en los ítems procesados. Por ejemplo, los sucesivos artículos fabricados en un proceso de producción no son exactamente idénticos, de la misma forma que los tiempos que tardan los usuarios en recibir un determinado servicio tampoco los son, incluso aunque el resultado final sea satisfactorio. Se considera que este tipo de variación, llamada *variación aleatoria*, es inherente al sistema. Sin embargo, existe otro tipo de variación que aparece ocasionalmente. Esta

variación, lejos de ser inherente al sistema se debe a causas concretas y habitualmente tiene un efecto negativo sobre la calidad de las operaciones industriales. Por ejemplo, en el contexto de producción, esta última variación puede ser debida a un fallo en la maquinaria empleada, o a una falta de calidad en la materia prima utilizada, o a un *software* incorrecto, o a un error humano, o a cualquiera otra causa posible. Cuando la única variación presente se debe al azar y no es, pues, asignable a ninguna causa se dice que el proceso está *bajo control*. Un problema clave en el control de la calidad consiste en reconocer cuándo un determinado proceso se encuentra bajo control o fuera de él.

En este capítulo se estudiarán los gráficos de control, que se pueden utilizar para reconocer si un proceso particular está fuera de control. Los tipos de gráficos de control que se considerarán vienen determinados por dos valores, llamados *límite de control superior* (*UCL*, del inglés *upper control limit*) y *límite de control inferior* (*LCL*, iniciales de *lower control limit*). Para utilizar esos gráficos, primero los datos generados en el proceso industrial se dividen en subgrupos. Después se calculan las medias de los subgrupos y, cuando una de ellas no cae entre los límites de control superior e inferior, se concluye que el proceso está fuera de control.

En la sección 15.2 se supondrá que los sucesivos elementos procesados presentan una característica medible –relacionada con su nivel de calidad, tanto si se trata de artículos producidos o de servicios realizados– cuya media y varianza son conocidas cuando el proceso está operando bajo control. Se verá cómo se pueden construir gráficos de control que permitan detectar un cambio en la media de la distribución bajo control. En la sección 15.3 se construirán gráficos de control en situaciones en las que cada ítem, en vez de presentar una característica medible, se clasifica bien como satisfactorio o bien como insatisfactorio. En las secciones 15.4 y 15.5 se introducirán dos tipos de gráficos de control que son particularmente eficientes para detectar pequeños deslizamientos en el valor medio de un proceso: los *gráficos de control de medias móviles ponderadas exponencialmente* y los *gráficos de control de sumas acumuladas*, respectivamente.

15.2 Gráfico de control de $\overline{X}$ para detectar un deslizamiento en la media

Supongamos que cuando un sistema industrial está bajo control, los sucesivos ítems procesados pueden tomar valores que son variables aleatorias independientes con media μ y varianza σ^2. Sin embargo supongamos que, por circunstancias impredecibles, el proceso puede irse de fuera de control y, como resultado, comienza a procesar ítems cuyos valores provienen de una distribución diferente. Nos interesará ser capaces de reconocer cuándo ocurre este hecho para poder parar el proceso, averiguar cuál es el fallo y arreglarlo.

Denotemos por $X_1, \ldots, X_n$ a los valores de los sucesivos ítems procesados por el sistema. Para intentar detectar cuándo el proceso está fuera de control será conveniente primero dividir los datos en subgrupos de tamaño fijo –digamos n–. Entre otras causas, este valor n se deberá elegir de forma que consigamos una cierta uniformidad de los valores de los datos dentro de los subgrupos individuales. Esto es, se intentará elegir n de modo que sea razonable que, cuando se produzca un deslizamiento en la distribución, se detecte entre y no dentro de los subgrupos. Así pues, en la práctica, habitualmente n se elige de modo que todos los datos dentro de un subgrupo se refieran a ítems procesados el mismo día, con los mismos parámetros, en las mismas condiciones, etc.

Denotemos como $\overline{X}_i$, $i = 1, 2, \ldots$, a la media del subgrupo i-ésimo. Puesto que cuando el proceso está bajo control todos los datos son normales con media μ y varianza σ^2, se tendrá que $\overline{X}_i$, la media de n de ellos, se distribuye igualmente según una normal cuyas media y varianza vienen dadas, respectivamente, por

$$E[\overline{X}_i] = \mu$$

$$\text{Var}(\overline{X}_i) = \frac{\sigma^2}{n}$$

Por consiguiente, cuando el proceso está bajo control,

$$Z = \frac{\overline{X}_i - \mu}{\sqrt{\sigma^2/n}}$$

es una variable aleatoria normal estándar. Esto es, si el proceso se mantiene bajo control durante el proceso de los elementos del grupo i, $\sqrt{n}\,(\overline{X}_i - \mu)/\sigma$ sigue una distribución normal estándar. Ahora bien, una variable aleatoria normal estándar, Z, está casi con seguridad comprendida entre -3 y $+3$. De hecho, se puede ver en la tabla 6.1 que $P\{-3 < Z < 3\} = 0{,}9973$. Por consiguiente, si el proceso de los ítems del subgrupo i se mantiene bajo control, se puede esperar casi con certeza que

$$-3 < \frac{\sqrt{n}(\overline{X}_i - \mu)}{\sigma} < 3$$

o, equivalentemente, que

$$\mu - \frac{3\sigma}{\sqrt{n}} < \overline{X}_i < \mu + \frac{3\sigma}{\sqrt{n}}$$

Los valores

$$\text{LCL} = \mu - \frac{3\sigma}{\sqrt{n}}$$

$$\text{UCL} = \mu + \frac{3\sigma}{\sqrt{n}}$$

se denominan *límite de control inferior* y *límite de control superior*, respectivamente.

El gráfico de control de $\overline{X}$, que fue diseñado originariamente para detectar un cambio en el valor medio de los ítems procesados, se obtiene si se representan gráficamente las medias sucesivas $\overline{X}_i$ de los subgrupos y si se declara que el proceso está fuera de control tan pronto como aparezca una media $\overline{X}_i$ que no caiga entre los límites LCL y UCL (Véase la figura 15.1).

Puesto que con los gráficos de control de $\overline{X}$ se declarará que un proceso está fuera de control sólo cuando la media de un subgrupo caiga fuera de los límites de control, es importante que los subgrupos se elijan de modo que sea altamente probable que cualquier deslizamiento que pueda ocurrir en la distribución sea detectado entre los subgrupos. Esto es así porque es más fácil detectar un deslizamiento en un subgrupo que tenga todos sus valores fuera de control que en un subgrupo en el que sólo algunos de sus valores estén fuera de control.

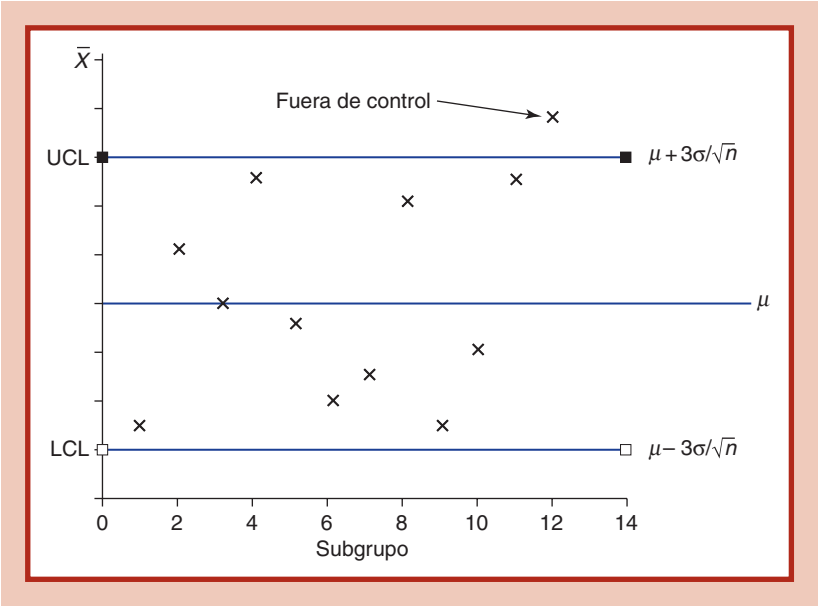


Figura 15.1 Gráfico de control de $\bar{X}$; n = tamaño del subgrupo.

Ejemplo 15.1 El tiempo que emplea una compañía de servicios informáticos en instalar un disco duro junto con un determinado programa informático es una variable aleatoria con media 25 minutos y desviación típica 6 minutos. La compañía dispone de dos empleados que trabajan en esta operación. Para monitorizar la eficiencia de estos empleados, la compañía ha representado gráficamente los tiempos medios sucesivos que han empleado en completar cuatro trabajos. Los subgrupos con número impar se refieren al primer empleado y los subgrupos con número par se refieren al segundo. Supongamos que las 20 primeras medias sucesivas son las siguientes (véase el gráfico de la figura 15.2):

Subgrupo	X	Subgrupo	X	Subgrupo	X	Subgrupo	X
1	23,6	6	24,6	11	29,4	16	32,8
2	20,8	7	22,6	12	27,8	17	23,3
3	25,5	8	24,4	13	26,8	18	30,5
4	26,2	9	24,7	14	27,2	29	25,3
5	23,3	10	26,0	15	24,0	20	34,1

¿Qué conclusiones se pueden extraer?

Solución Dado que el tamaño de los subgrupos es 4 y que las medias sucesivas tienen media $\mu = 25$ y desviación típica $\sigma = 6$ cuando el proceso está bajo control, se tiene que los límites de control vendrán dados por

$$\text{LCL} = 25 - \frac{18}{\sqrt{4}} = 16 \qquad \text{UCL} = 25 + \frac{18}{\sqrt{4}} = 34$$

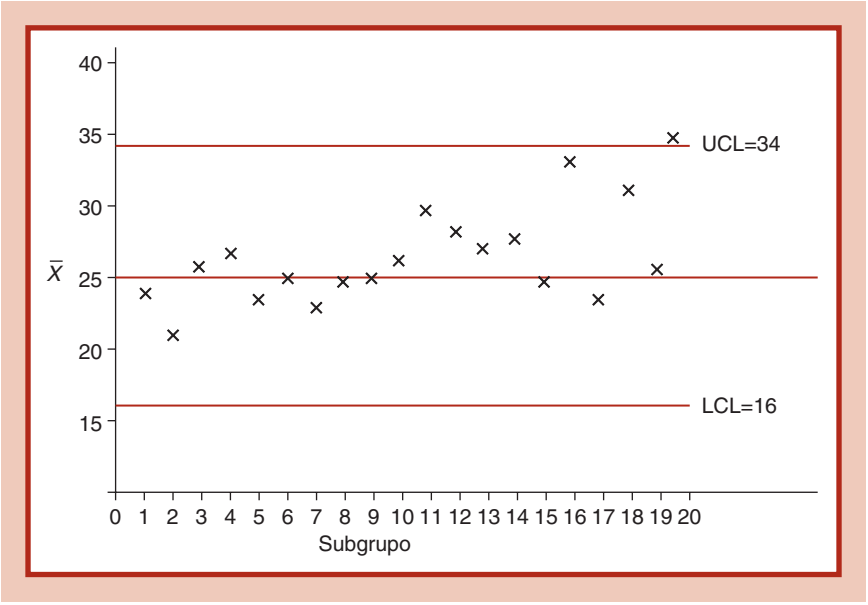


Figura 15.2 Gráfico de control para los datos del ejemplo 15.1.

Puesto que la media del subgrupo 20 es mayor que UCL, parece que el sistema ha dejado de estar bajo control. De hecho, dado que las seis últimas medias de los subgrupos pares son mayores que la media bajo control que es igual a 25 (entre las que las tres últimas son significativamente mayores) parece probable que el segundo empleado haya estado fuera de control durante un tiempo. ■

Se ha asumido hasta ahora que, cuando el proceso está bajo control, la distribución subyacente de la característica medible de los ítems procesados es normal; de hecho, esto es lo que ocurre habitualmente en los procesos de producción. Sin embargo, aún cuando no sea así, la distribución de las medias de los subgrupos se aproximará a la normal, como consecuencia del teorema central del límite; así pues, también en este caso continuará siendo muy improbable que las medias de los subgrupos difieran de su media por encima de 3 veces su desviación típica. Como resultado, cuando se utilizan las medias de los subgrupos no es necesario que se conozca exactamente la distribución bajo control, basta con conocer su media y su varianza. Ésta es la razón principal por la que se utilizan las medias de los subgrupos en lugar de los datos individuales. Los tamaños habituales de los subgrupos son 4, 5 y 6. La razón estriba en que si se usan tamaños de subgrupo más pequeños, la normalidad aproximada de las medias de los subgrupos podría ser cuestionable. Por otro lado, dado que no se puede detectar que el sistema esté bajo control hasta que se hayan procesado todos los ítems de al menos un subgrupo, no resulta aconsejable que el tamaño de los subgrupos sea excesivamente grande.

Los gráficos de control de $\overline{X}$ fueron propuestos inicialmente por Walter Shewart; por ello, se conocen también con el nombre de *gráficos de control de Shewart.* En esencia, estos gráficos equivalen a realizar para cada subgrupo el contraste estadístico de la hipótesis

$$H_0\text{: media} = \mu \qquad \text{frente a la alternativa} \qquad H_1\text{: media} \neq \mu,$$

al nivel de significación $\alpha = P\{|Z| > 3\} = 0{,}0027$. En cuanto la hipótesis nula resulta rechazada se declara que el proceso se encuentra fuera de control.

Aunque habitualmente los gráficos de control se aplican dentro del contexto de los procesos de producción, también se pueden utilizar en una gran variedad de situaciones, tal como se muestra en el siguiente ejemplo.

Ejemplo 15.2 Consideremos una pequeña tienda de alquiler de vídeo-películas cuyos alquileres diarios de lunes a jueves tienen media 52 y desviación típica 10. Si el número de películas alquiladas diariamente (de lunes a jueves) durante la semana pasada fueron

$$32, 38, 28, 30$$

¿se puede concluir que se ha producido un cambio en la distribución de los alquileres?

Solución Supongamos que los subgrupos están formados por los alquileres que se realizan en esos cuatro días de la semana. Puesto que la media y la desviación típica bajo control son $\mu = 52$ y $\sigma = 10$, respectivamente, los límites de control son

$$\text{LCL} = 52 - \frac{3(10)}{\sqrt{4}} = 37$$

$$\text{UCL} = 52 + \frac{3(10)}{\sqrt{4}} = 67$$

Puesto que la media del último subgrupo es $(32 + 38 + 28 + 30)/4 = 32$, que se encuentra fuera de los anteriores límites, se puede mantener que el número medio de alquileres diarios ha dejado de ser igual a 52.

En esta situación, el encargado del almacén debería intentar descubrir (1) la causa del cambio en la media y (2) si este cambio es temporal o permanente. Por ejemplo, el encargado podría averiguar que se han emitido durante esos días de la semana pasada ciertos programas de televisión particularmente interesantes, tales como los Campeonatos Mundiales, o las Olimpiadas, o un acto político relevante, que justifiquen la idea de que el cambio se produjo sólo a corto plazo. O también se podría descubrir que el cambio en media pudo deberse a la aparición de un nuevo competidor en la vecindad, en cuyo caso podría ser más permanente. ■

En ocasiones no se miden y se anotan todas las unidades producidas sino solamente un subconjunto aleatorio de ellas. Cuando esto ocurre es natural que los artículos que componen cada subgrupo hayan sido producidos más o menos en un corto periodo de tiempo.

Problemas

1. Cuando está bajo control, un proceso fabrica ítems que tienen media 100 y desviación típica 10. Determine los límites de control superior e inferior para las medias de los subgrupos cuando el tamaño del subgrupo es:

   (a) 4 (b) 5 (c) 6 (d) 10

2. Cuando un proceso está funcionando correctamente produce ítems que tienen media 35 y desviación típica 4. Para monitorizar este proceso se muestrean subgrupos de tamaño 4. Si en la siguiente tabla se muestran las medias de los primeros 20 subgrupos, ¿parece que el proceso ha estado bajo control durante todo el tiempo en que se produjeron?

Subgrupo	$\overline{X}$	Subgrupo	$\overline{X}$
1	31,2	11	36,4
2	38,4	12	31,1
3	35,0	13	32,3
4	33,3	14	37,8
5	34,7	15	36,6
6	31,1	16	40,4
7	35,8	17	41,2
8	34,4	18	35,9
9	37,1	19	40,4
10	34,2	20	32,5

3. Cuando un determinado proceso de fabricación está bajo control se producen cables cuyo diámetro medio es 80 con una desviación típica de 10 (en unidades de 1/10 000 pulgadas). Los datos siguientes representan las medias muestrales de subgrupos de tamaño 5:

   85, 88, 90, 77, 79, 83, 90, 75, 94, 80, 84, 86, 88

   ¿Parece que el proceso ha estado bajo control?

4. Se quiere disponer de un gráfico de control para analizar los tiempos que tardan los trabajadores en llevar a cabo una determinada tarea. El tiempo que debería costarles sigue una normal con una media de 26 minutos y una desviación típica de 4,2 minutos. Los siguientes datos representan los valores de $\overline{X}$ para 10 subgrupos de tamaño 4:

   28,2, 28,4, 31,1, 27,3, 33,2, 31,4, 27,9, 30,4, 31,3, 30,4

   (a) Determine los límites de control superior e inferior.

   (b) ¿Parece que el proceso haya estado bajo control?

5. Cuando un proceso de soldadura de juntas funciona correctamente, la distancia de la soldadura al centro de la junta se distribuye según una normal con media 0 y desviación típica 0,05 pulgadas. Si los siguientes valores muestran las distancias medias entre las soldaduras y los centros de las juntas para 8 subgrupos de tamaño 5, ¿parece que el proceso haya estado bajo control cuando se llevaron a cabo las mediciones?

Subgrupo	Distancia media
1	0,0023
2	−0,0012
3	−0,0015
4	0,0031
5	0,0038
6	0,0051
7	0,0022
8	−0,0033

6. Antes de 1995, el número de asesinatos cometidos anualmente en Estados Unidos por cada 100 000 habitantes seguía una normal con media 9,0 y desviación típica 1,1. Los siguientes datos muestran esas tasas para los años comprendidos entre 1995 y 2002:

8,4, 7,4, 6,8, 6,3, 5,7, 5,5, 5,6, 5,4

¿Se puede concluir que la tasa de asesinatos ha cambiado con respecto a su valor histórico? Utilice subgrupos de tamaño 2.

15.2.1 Cuando la media y la varianza son desconocidas

Cuando se está empezando a construir un gráfico de control y no se dispone de datos históricos fiables, se deben estimar la media μ y la desviación típica σ. Para hacerlo, se emplean k de los subgrupos haciendo, cuando sea posible, que $k \geq 20$ y $nk \geq 100$, donde n representa el tamaño de los subgrupos. Si $\overline{X}_i$ es la media del grupo i, se puede estimar μ mediante $\overline{\overline{X}}$, la media de todas las medias de los subgrupos. Esto es,

$$\overline{\overline{X}} = \frac{\overline{X}_1 + \cdots + \overline{X}_k}{k}$$

Dado que $\overline{\overline{X}}$ es la media de todos los nk valores de datos, éste es el estimador "natural" de la media poblacional μ.

Para estimar σ, sea S_i la desviación típica muestral de los datos del subgrupo i, $i = 1, \ldots, k$, y sea $\overline{S}$ la media de las desviaciones típicas de todos los subgrupos. Esto es,

$$\overline{S} = \frac{S_1 + \cdots + S_k}{k}$$

Puesto que el valor esperado de $\overline{S}$ no coincide con σ, se ha de dividir por una constante $c(n)$ que depende del tamaño n de los subgrupos, para conseguir un estimador insesgado de σ. Esto es, se utiliza el estimador $\overline{S}/c(n)$, verificándose que

$$E[\overline{S}/c(n)] = \sigma$$

En la tabla siguiente se muestran los valores de $c(n)$, cuando n varía entre 3 y 9.

Valores de $c(n)$
$c(3) = 0{,}8862266$
$c(4) = 0{,}9213181$
$c(5) = 0{,}9399851$
$c(6) = 0{,}9515332$
$c(7) = 0{,}9593684$
$c(8) = 0{,}9650309$
$c(9) = 0{,}9693103$

Los anteriores estimadores de μ y σ usan los k subgrupos y son aceptables siempre que el sistema se haya mantenido bajo control a lo largo del procesamiento de todos ellos. Para comprobar esto primero se calculan los siguientes límites de control basados en estos estimadores:

$$\text{LCL} = \overline{\overline{X}} - \frac{3\overline{S}}{\sqrt{n}c(n)}$$

$$\text{UCL} = \overline{\overline{X}} + \frac{3\overline{S}}{\sqrt{n}c(n)}$$

Después se comprueba que las medias de los k subgrupos caen dentro de estos límites. Si alguna de ellas cayera fuera se debe decidir si es razonable suponer que el sistema ha estado temporalmente fuera de control mientras se procesaban los ítems del subgrupo. Si es así, o si se ha descubierto la razón por la que el proceso estuvo fuera de control se han de eliminar los subgrupos con problemas, sustituirlos por otros cuyo procesamiento haya estado bajo control, y recalcular los estimadores de μ y σ. Se continúa con esto hasta que las medias de todos los subgrupos se mantengan dentro de los límites de control estimados. Claramente, cuantas más medias de los subgrupos caigan fuera de los límites de control, mayor será la evidencia de que se ha perdido el control.

Ejemplo 15.3 Consideremos que una nueva factoría que produce acondicionadores de aire para automóviles que liberan solamente cantidades mínimas de clorofluorocarbonos dañinos. Una vez producido un acondicionador cualquiera, se analiza qué cantidad de clorofluorocarbonos, medida en unidades adecuadas, libera durante una hora de funcionamiento. Los datos siguientes muestran las medias muestrales y las desviaciones típicas

muestrales correspondientes a 50 acondicionadores de aire divididos en 10 subgrupos de tamaño 5:

i	$\bar{X}_i$	S_i
1	30,1	1,22
2	29,7	1,40
3	31,2	0,81
4	29,9	1,10
5	30,3	0,93
6	30,2	0,82
7	31,0	1,54
8	31,4	1,58
9	30,9	1,26
10	32,0	1,60

Los valores de los estimadores de μ y σ son

$$\bar{\bar{X}} = 30\,670, \qquad \frac{\bar{S}}{c(5)} = \frac{1{,}226}{0{,}9399851} = 1{,}304$$

Puesto que $3(1{,}304)/\sqrt{5} = 1{,}750 = 1{,}750$, se tiene que los límites de control estimados son

$$\text{LCL} = 30{,}670 - 1{,}750 = 28{,}920$$
$$\text{UCL} = 30{,}670 + 1{,}750 = 32{,}420$$

Como todas las medias de los subgrupos caen entre los límites se puede suponer que el proceso ha estado bajo control con $\mu = 30{,}670$ y $\sigma = 1{,}304$.

Supongamos ahora que se exige que al menos un 99% de los acondicionadores no liberen más de 33,40 unidades de clorofluorocarbonos por hora. Si se asume que el proceso de producción se mantiene bajo control con los anteriores valores estimados de la media y de la desviación típica, ¿se cumplirá tal exigencia? Para contestar a esta pregunta, observe que si X, la cantidad de clorofluorocarbono liberado por un acondicionador de aire en una hora es una variable aleatoria normal con media 30,670 y desviación típica 1,304; por consiguiente,

$$\begin{aligned} P\{X > 33{,}40\} &= P\left\{\frac{X - 30{,}670}{1{,}304} > \frac{33{,}40 - 30{,}670}{1{,}304}\right\} \\ &= P\{Z > 2{,}094\} \\ &= 0{,}018 \end{aligned}$$

De aquí se desprende que un 1,8% de los acondicionadores liberan en una hora más de 33,4 unidades. Así pues, no se cumple la exigencia de que al menos un 99% de ellos satisfaga esta condición. ■

Observación: El estimador $\overline{\overline{X}}$ coincide con la media de todos los nk valores de datos; es, pues, el estimador obvio de la media μ. Sin embargo, puede no resultar evidente por qué no se utiliza la desviación típica muestral de todos los nk valores de datos, es decir, $\sqrt{S^2}$, siendo

$$S^2 = \frac{\sum_{i=1}^{nk}(X_i - \overline{\overline{X}})^2}{nk - 1}$$

para estimar σ. No se hace esto porque el sistema puede no haber estado bajo control durante el proceso de los primeros k subgrupos y, por esta razón, este último estimador puede diferir mucho de verdadero valor de σ bajo control. Por otro lado, incluso aunque se hubiera producido un cambio en el valor medio durante el procesamiento de los k subgrupos, si se supone que la desviación típica ha permanecido constante, el estimador $\overline{S}/c(n)$ continuará siendo un buen estimador de σ, dado que sólo requiere que los valores de datos en cada subgrupo tengan la misma media (que puede diferir de un subgrupo a otro).

15.2.2 Gráficos de control para S

Los gráficos de control de $\overline{X}$ están diseñados para recoger cualquier cambio en la media poblacional. En aquellos casos en los que se estuviera interesado en reconocer los cambios producidos en la desviación típica poblacional se pueden utilizar los gráficos de control para S.

Como en las secciones anteriores, supongamos que los ítems producidos en un proceso que se encuentre bajo control tienen media μ y desviación típica σ. Si S_i es la desviación típica muestral para el subgrupo i, se tiene que, tal como se observó en la sección 15.2.1,

$$E\left[\frac{S_i}{c(n)}\right] = \sigma$$

Esto implica que

$$E[S_i] = c(n)\sigma$$

siendo n el tamaño de los subgrupos. Adicionalmente, con la identidad

$$\text{Var}(S_i) = E[S_i^2] - (E[S_i])^2$$

se obtiene de lo anterior y del hecho de que el valor esperado de la varianza muestral es la varianza poblacional, que

$$\begin{aligned}\text{Var}(S_i) &= \sigma^2 - c^2(n)\sigma^2 \\ &= \sigma^2[1 - c^2(n)]\end{aligned}$$

Esto es, si se supone que el proceso se encuentra bajo control, S_i es una variable aleatoria con media $c(n)\sigma$ y desviación típica $\sigma\sqrt{1 - c^2(n)}$. De aquí se sigue que, dado que una variable aleatoria es muy improbable que tome un valor que difiera de la media en más de tres veces su desviación típica, parece razonable fijar los límites de control del gráfico de control para S como sigue:

$$\text{LCL} = c(n)\sigma - 3\sigma\sqrt{1 - c^2(n)}$$

$$\text{UCL} = c(n)\sigma + 3\sigma\sqrt{1 - c^2(n)}$$

Se deberían representar gráficamente los sucesivos valores de S_i para estar seguros de que caen dentro de los límites control. Cuando un valor caiga fuera de ellos, el proceso se debería parar y ser declarado fuera de control.

Cuando se está empezando a diseñar un gráfico de control para S y se supone que σ es desconocido, se podría estimar mediante $\bar{S}/c(n)$, siendo $\bar{S}$ la media de las desviaciones típicas de los k subgrupos. Teniendo en cuenta lo anterior, los límites de control estimados coinciden con

$$\text{LCL} = \bar{S}\left[1 - 3\sqrt{\frac{1}{c^2(n)} - 1}\right]$$

$$\text{UCL} = \bar{S}\left[1 + 3\sqrt{\frac{1}{c^2(n)} - 1}\right]$$

Al igual que cuando se empieza a construir un gráfico de control para $\bar{X}$, se debería comprobar que las k desviaciones típicas de los subgrupos caen dentro de los anteriores límites de control. Se habrían de descargar aquellos que caen fuera, y se debería recalcular el estimador de σ (posiblemente con datos adicionales).

Ejemplo 15.4 Los siguientes datos de subgrupos representan las medias y las desviaciones típicas (en minutos) para 20 subgrupos de tamaño 5 correspondientes a un nuevo proceso de producción de dardos de acero. Los datos se refieren a los tiempos de producción.

Subgrupo	$\bar{X}$	S	Subgrupo	$\bar{X}$	S	Subgrupo	$\bar{X}$	S
1	35,1	4,2	8	38,4	5,1	15	43,2	3,5
2	33,2	4,4	9	35,7	3,8	16	41,3	8,2
3	31,7	2,5	10	27,2	6,2	17	35,7	8,1
4	35,4	3,2	11	38,1	4,2	18	36,3	4,2
5	34,5	2,6	12	37,6	3,9	19	35,4	4,1
6	36,4	4,5	13	38,8	3,2	20	34,6	3,7
7	35,9	3,4	14	34,3	4,0			

Puesto que

$$\overline{\overline{X}} = 35{,}94, \qquad \overline{S} = 4{,}35, \qquad c(5) = 0{,}9400$$

se obtienen los siguientes límites de control superiores e inferiores para $\overline{X}$ y S, respectivamente:

$$\text{LCL}(\overline{X}) = 35{,}94 - \frac{3(4{,}35)}{0{,}94\sqrt{5}} = 29{,}731$$

$$\text{UCL}(\overline{X}) = 35{,}94 + \frac{3(4{,}35)}{0{,}94\sqrt{5}} = 42{,}149$$

$$\text{LCL}(S) = 4{,}35\left[1 - 3\sqrt{\frac{1}{(0{,}94)^2 - 1}}\right] = -0{,}386$$

$$\text{UCL}(S) = 4{,}35\left[1 + 3\sqrt{\frac{1}{(0{,}94)^2 - 1}}\right] = 9{,}087$$

Las figuras 15.3 y 15.4 muestran los gráficos de control para $\overline{X}$ y S, con los anteriores límites de control. Puesto que las medias 10 y 15 caen fuera de los límites de control, estos subgrupos se deberían eliminar y, tras ello, se tendrían que recalcular los límites. Se deja como ejercicio la realización de los cálculos necesarios. ■

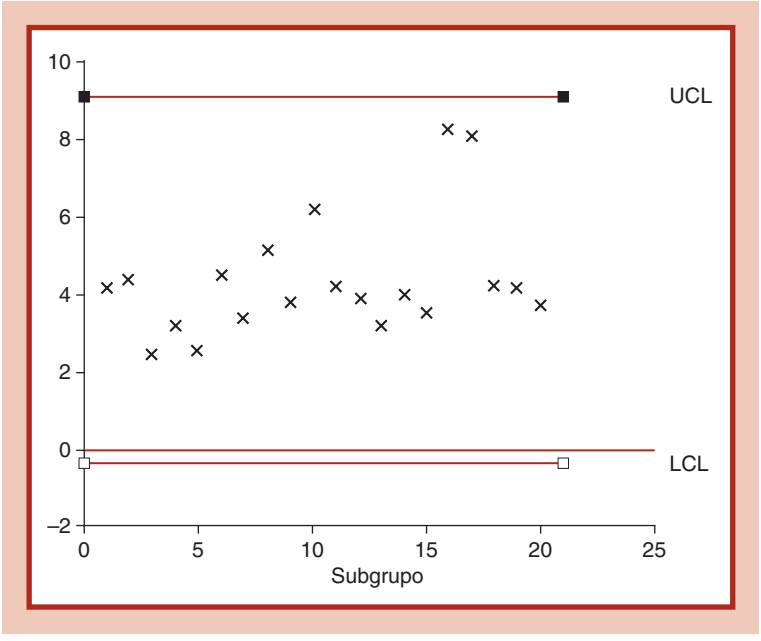


Figura 15.3 Gráfico de control para S.

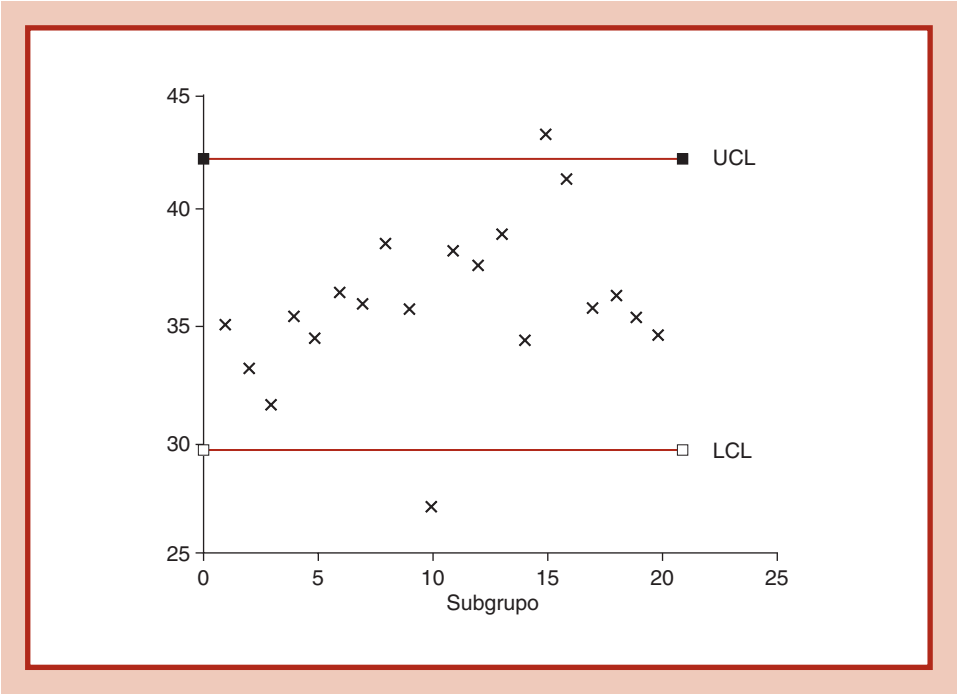


Figura 15.4 Gráfico de control para $\overline{X}$.

Problemas

1. Se han recogido datos de 150 ítems producidos mediante un nuevo proceso de producción. Los datos se han dividido en 25 subgrupos de tamaño 6 y, tras ello, se han determinado las medias muestrales y las desviaciones típicas muestrales de los distintos subgrupos. Supongamos que

$$\sum_{i=1}^{25} \overline{X}_i = 357{,}3, \qquad \sum_{i=1}^{25} S_i = 5{,}44$$

(a) Con los datos anteriores, obtenga los límites de control del gráfico de control para $\overline{X}$. Asuma para ello que las 25 medias de los subgrupos caen dentro de estos límites.

(b) Supongamos que se exige que los ítems producidos tengan una medida que caiga dentro del rango $14{,}3 \pm 0{,}50$. Suponiendo que la media y la varianza son iguales a los estimadores obtenidos a partir de los anteriores datos cuando el proceso está bajo control, ¿qué porcentaje de ítems tienen valores comprendidos dentro del rango citado?

2. Los datos siguientes muestran los valores de $\overline{X}$ y S para 20 subgrupos de tamaño 5:

Subgrupo	$\overline{X}$	S	Subgrupo	$\overline{X}$	S
1	33,8	5,1	11	29,7	5,1
2	37,2	5,4	12	31,6	5,3
3	40,4	6,1	13	38,4	5,8
4	39,3	5,5	14	40,2	6,4
5	41,1	5,2	15	35,6	4,8
6	40,4	4,8	16	36,4	4,6
7	35,0	5,0	17	37,2	6,1
8	36,1	4,1	18	31,3	5,7
9	38,2	7,3	19	33,6	5,5
10	32,4	6,6	20	36,7	4,2

(a) Determine los límites de control tentativos del gráfico de control para $\overline{X}$.

(b) ¿Parece que el proceso se ha mantenido bajo control a lo largo del proceso de los subgrupos?

(c) Estime el porcentaje de ítems producidos que tienen valores comprendidos entre 25 y 45.

3. Los datos siguientes muestran las sucesivas medias y desviaciones típicas de los subgrupos de datos, referidos a una característica eléctrica (medida en decibelios) de láminas de cerámica. Los subgrupos son de tamaño 4.

$\overline{X}$: 16,1, 15,7, 16,6, 16,0, 14,7, 15,8, 16,4, 14,5, 15,8, 17,2

S: 2,7, 2,9, 2,2, 1,0, 1,3, 2,6, 3,1, 2,5, 5,3, 4,4

(a) Utilice los datos anteriores para estimar la media y la desviación típica de la población.

(b) Aparentemente, ¿ha estado el proceso bajo control?

4. Los datos históricos indican que el tiempo que se tarda en llevar a cabo un determinado trabajo tiene una media de 26 minutos y una desviación típica de 8,3 minutos.

(a) Determine los límites de control del gráfico de control para $\overline{X}$ utilizando subgrupos de tamaño 5.

(b) Determine los límites de control del gráfico de control para S usando subgrupos de tamaño 5.

5. Los siguientes datos se refieren a las cantidades en las que los diámetros de alambre, medidos en unidades de 0,001 pulgadas, sobrepasan un determinado valor.

Subgrupo	Valores de datos				
1	2,5	0,5	2,0	−1,2	1,4
2	0,2	0,3	0,5	1,1	1,5
3	1,5	1,3	1,2	−1,0	0,7
4	0,2	0,5	−2,0	0,0	−1,3
5	−0,2	0,1	0,3	−0,6	0,5
6	1,1	−0,5	0,6	0,5	0,2
7	1,1	−1,0	−1,2	1,3	0,1
8	0,2	−1,5	−0,5	1,5	0,3
9	−2,0	−1,5	1,6	1,4	0,1
10	−0,5	3,2	−0,1	−1,0	−1,5
11	0,1	1,5	−0,2	0,3	2,1
12	0,0	−2,0	−0,5	0,6	−0,5
13	−1,0	−0,5	−0,5	−1,0	0,2
14	0,5	1,3	−1,2	−0,5	−2,7
15	1,1	0,8	1,5	−1,5	1,2

(a) Obtenga los límites de control tentativos de los gráficos de control para $\overline{X}$ y S.

(b) Aparentemente, ¿ha estado el proceso siempre bajo control?

(c) Si la respuesta al apartado (b) es no, calcule los límites de control revisados.

6. Los siguientes datos muestran los valores $\overline{X}$ y S para los primeros 10 subgrupos de tamaño 5 fabricados en una nueva cadena de producción:

$\overline{X}$: 20,2, 28,4, 31,1, 27,3, 33,2, 31,4, 27,9, 30,4, 31,3, 30,4

S: 7,2, 2,8, 3,4, 4,1, 4,0, 3,3, 4,5, 3,0, 2,7, 2,1

(a) Utilice los datos anteriores para calcular los límites de control tentativos de los gráficos de control para $\overline{X}$ y S, respectivamente.

(b) ¿Parece que el proceso ha estado siempre bajo control?

(c) Si no se dispone de datos adicionales, ¿qué límites de control se deberían utilizar para los datos futuros?

7. Complete el ejemplo 15.4.

8. Para el problema 1, determine los límites de control tentativos del gráfico de control para S.
9. Para el problema 2, calcule los límites de control tentativos del gráfico de control para S.

15.3 Gráficos de control para la fracción de defectos

Los gráficos de control para $\overline{X}$ se utilizan cuando la variable de medida toma valores continuos dentro de una determinada región. Sin embargo, en ciertas ocasiones los ítems procesados sólo se pueden clasificar como aceptables o inaceptables. Por ejemplo, un artículo producido puede ser catalogado como defectuoso o no; o un cliente puede considerar un determinado servicio recibido como aceptable o no. En este apartado, se verá cómo se pueden construir gráficos de control en esos casos.

Supongamos que, cuando el sistema se encuentra bajo control, cada ítem procesado puede independientemente ser defectuoso con probabilidad p. Si se denota por X el número de ítems defectuosos en un subgrupo de tamaño n, se tendrá que, asumiendo que el sistema está bajo control, X será una variable aleatoria binomial de parámetros n y p; así pues,

$$E[X] = np$$

$$\text{Var}(X) = np(1 - p)$$

De aquí se desprende que, cuando el sistema se encuentra bajo control, el número de defectos existentes en un subgrupo de tamaño n deberá, con una alta probabilidad, estar comprendido entre los límites de control inferior y superior

$$\text{LCL} = np - 3\sqrt{np(1 - p)}$$

$$\text{UCL} = np + 3\sqrt{np(1 - p)}$$

Por lo general, el tamaño n de los subgrupos suele ser mucho mayor que los valores habitualmente empleados para construir los gráficos de control para $\overline{X}$, que suelen estar comprendidos entre 4 y 10. Esto se debe principalmente a que, si p es pequeño (tal como ocurre habitualmente) y n no es lo suficientemente grande, la mayor parte de los subgrupos tendrán 0 defectos, incluso aunque el proceso se encuentre fuera de control; por consiguiente, serán más difíciles de detectar las situaciones de fuera de control que si se eligieran valores de n para los que el producto np no sea demasiado pequeño. Una segunda razón a favor de la utilización de valores grandes de n se basa en el hecho de que, cuando np tiene un valor moderado, X seguirá aproximadamente una distribución normal; por lo que, cuando se esté bajo control, el estadístico de cada subgrupo caerá dentro de los límites de control con una probabilidad aproximadamente igual a $1 - 0{,}0027 = 0{,}9973$.

Ejemplo 15.5 Se han extraído muestras sucesivas de 200 tornillos de la producción de una máquina automática. Cada tornillo puede ser catalogado como aceptable o defec-

tuoso. Supongamos que, a partir de los datos históricos, se sabe que cada tornillo es independientemente defectuoso con probabilidad 0,07, cuando el proceso está bajo control. Si los datos siguientes muestran el número de tornillos defectuosos en 20 muestras, ¿se debería haber declarado el proceso fuera de control, en algún momento durante la recogida de las muestras?

Subgrupo	Defectuosos	Subgrupo	Defectuosos
1	23	11	4
2	22	12	13
3	12	13	17
4	13	14	5
5	15	15	9
6	11	16	5
7	25	17	19
8	16	18	7
9	23	19	22
10	14	20	17

Solución Puesto que $n = 200$ y $p = 0{,}07$, se tiene que

$$np = 14 \qquad 3\sqrt{np(1-p)} = 10{,}825$$

así pues

$$\text{LCL} = 14 - 10{,}825 = 3{,}175$$

$$\text{UCL} = 14 + 10{,}825 = 24{,}825$$

Dado que el número de defectos del subgrupo 7 cae fuera del intervalo comprendido entre LCL y UCL, el proceso debería haberse declarado fuera de control en ese punto. ■

***Observación*:** Tenga en cuenta que se intenta detectar cualquier cambio que ocurra en la calidad, incluso aunque el cambio suponga una mejora en la misma. Esto es, se declara que el proceso está fuera de control incluso cuando disminuye la probabilidad de que un ítem sea defectuoso. La razón estriba en que es importante reconocer cualquier cambio en la calidad, a mejor o a peor, de modo que seamos capaces de evaluar la causa que justifica el cambio producido. En otras palabras, si se produce una mejora en la calidad de los productos procesados es importante determinar la razón de esta mejora (¿qué se está haciendo bien?).

Problemas

1. Los datos siguientes representan el número de defectos de ensamblaje encontrados en muestras de tamaño 200.

Número de muestra	Número de defectos	Número de muestra	Número de defectos
1	7	11	4
2	3	12	10
3	2	13	0
4	6	14	8
5	9	15	3
6	4	16	6
7	3	17	2
8	3	18	1
9	2	19	6
10	5	20	10

   Supongamos que cuando el proceso está bajo control cada ensamblaje es defectuoso con probabilidad 0,03. ¿Parece que el proceso haya estado siempre bajo control?

2. Supongamos que cuando un proceso se encuentra bajo control cada ítem producido es defectuoso con probabilidad 0,04. Si en un gráfico de control se toman muestras diarias de tamaño 500, calcule los límites de control superior e inferior.

3. Históricamente, un 4% de los contenedores de fibra son defectuosos debido a la contaminación derivada de la cola de pegar. Los datos siguientes representan el número de contenedores defectuosos en muestras sucesivas de tamaño 100.

   3, 5, 1, 0, 4, 7, 8, 9, 5, 7, 1, 3, 0, 5, 3, 6, 4, 8, 3, 6

   (a) ¿Parece que el proceso haya estado bajo control?

   (b) ¿Cuáles son los límites de control?

15.4 Gráficos de control de medias móviles ponderadas exponencialmente

Aunque los gráficos de control de $\overline{X}$ resultan muy efectivos para detectar los cambios grandes que se producen temporalmente en la media, no lo son tanto para detectar los cambios pequeños que tengan una tendencia a persistir en el tiempo. Por ejemplo, consideremos un

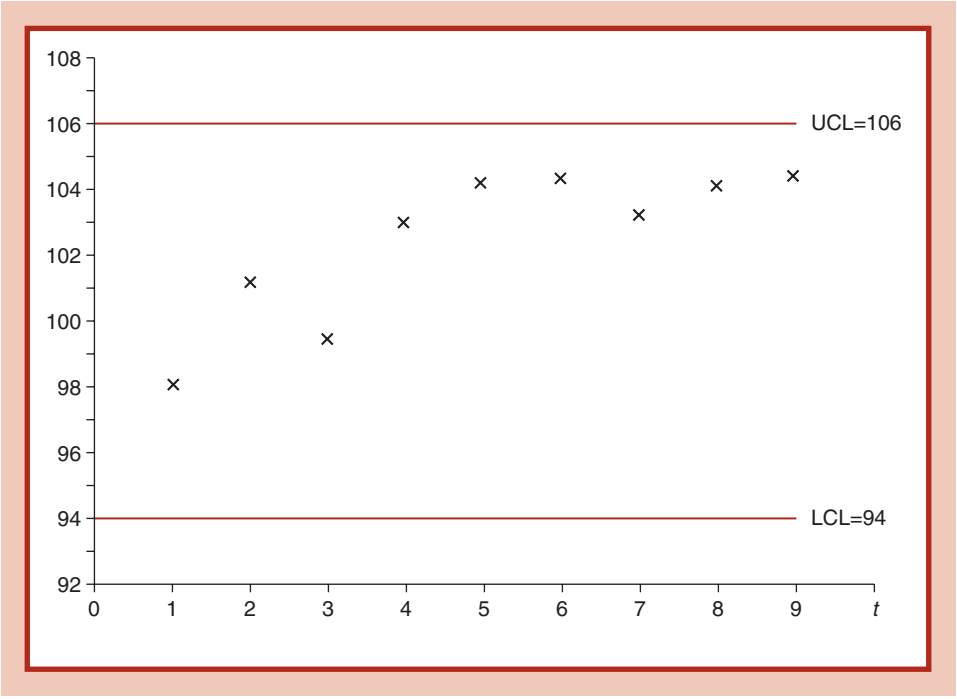


Figura 15.5 Medias de los sucesivos subgrupos.

proceso cuyas medias de los subgrupos tengan una media $\mu = 100$ y una desviación típica $\sigma/\sqrt{n} = 2$ cuando se encuentra bajo control. El gráfico de control para $\overline{X}$ tendría como límites de control 100 ± 6. Supongamos que las medias de los sucesivos subgrupos fueran (véase la figura 15.5):

98, 101,2, 99,4, 103, 104,2, 104,3, 103,2, 104,1, 104,4

Aunque es bastante obvio que el proceso está fuera de control (puesto que, entre otras causas, cuatro de las últimas cinco medias de subgrupos superan a μ en más de dos desviaciones típicas de los subgrupos), el gráfico de control de $\overline{X}$ no habría detectado este hecho. Lo anterior ilustra la debilidad fundamental de los gráficos de control de $\overline{X}$, en el sentido de que cada media de los subgrupos se trata aisladamente y no en relación con los valores de las medias de los subgrupos más próximos. Por esta razón es necesario que una media de subgrupo se encuentre muy separada de μ para que el proceso se declare fuera de control.

Se presentará a continuación un gráfico de control más sofisticado, conocido como *gráfico de medias móviles ponderadas exponencialmente*, que considera las medias de los subgrupos en relación con aquellas que están próximas. Para empezar, supongamos como antes que, si el proceso está bajo control, se producen ítems cuyos valores se distribuyen según una normal de media μ y desviación típica σ. Sea n el tamaño de los subgrupos y denotemos por $\overline{X}_i$ la media de los valores del subgrupo i, $i \geq 1$. Finalmente, sea β

una constante comprendida entre 0 y 1 y definamos la sucesión de valores $W_0, W_1, \ldots$ como sigue:

$$W_0 = \mu$$

$$W_t = \beta\bar{X}_t + (1 - \beta)W_{t-1}, \qquad t = 1, 2, \ldots$$

Esto es, el valor inicial de la sucesión W es μ y cada valor sucesivo es una media ponderada de la media del siguiente subgrupo y del valor anterior de W. La sucesión de valores $W_0, W_1, W_2, \ldots$ se conoce como la sucesión *de medias móviles ponderadas exponencialmente* (*EWMA*, del inglés e*xponentially weighted moving average*). (En ocasiones se llama también sucesión *de medias móviles ponderadas geométricamente.*) Se la denomina así porque se puede demostrar que W_t, el valor de la media móvil en el instante t, se puede expresar como

$$W_t = \beta\bar{X}_t + \beta(1 - \beta)\bar{X}_{t-1} + \beta(1 - \beta)^2\bar{X}_{t-2} + \cdots + \beta(1 - \beta)^{t-1}\bar{X}_1 + \beta(1 - \beta)^t\mu$$

En otras palabras, W_t es una media ponderada de todas las medias de los subgrupos obtenidas hasta el instante t, en la que se otorga un peso β a la media más reciente, los sucesivos pesos van decreciendo al multiplicar el peso anterior por el factor $1 - \beta$ y, finalmente, se otorga el peso $\beta(1 - \beta)^t$ a la media bajo control μ.

Cuanto menor sea el valor elegido de β, más iguales serán los sucesivos pesos. Por ejemplo, si se toma $\beta = 0{,}2$, los pesos sucesivos son 0,2, 0,16, 0,128, 0,1024, 0,08192, y así sucesivamente; mientras que si $\beta = 0{,}9$, los pesos sucesivos son 0,9, 0,09, 0,0009, etcétera.

Puesto que W_t se puede expresar como una suma de variables aleatorias normales e independientes, su distribución continúa siendo normal. Su valor esperado es

$$E[W_t] = \mu$$

y, para valores moderadamente altos de t, su desviación típica viene dada aproximadamente por

$$\mathrm{SD}(W_t) = \sqrt{\frac{\beta}{2 - \beta}}\,\frac{\sigma}{\sqrt{n}}$$

El gráfico de control en el que sucesivamente se representa gráficamente W_t y se declara el proceso fuera de control tan pronto como algún W_t caiga fuera de los límites de control

$$\mathrm{LCL} = \mu - 3\sqrt{\frac{\beta}{2 - \beta}}\,\frac{\sigma}{\sqrt{n}}$$

$$\mathrm{UCL} = \mu + 3\sqrt{\frac{\beta}{2 - \beta}}\,\frac{\sigma}{\sqrt{n}}$$

se denomina *gráfico de control estándar de medias móviles con ponderación exponencial* y factor de ponderación β.

Ejemplo 15.6 Un empresa de reparación de equipos electrónicos envía a uno de sus técnicos a aquellos hogares que requieren sus servicios. Tras recibir un encargo, la empresa envía a un empleado con la instrucción de llamar por teléfono en cuanto se termine el trabajo. Los datos históricos muestran que el tiempo transcurrido desde que sale el técnico hasta que llama por teléfono es una variable aleatoria normal con media 62 minutos y desviación típica 24 minutos. Para ser conscientes de los cambios que se pudieran producir en esta distribución, la empresa de reparación hace uso de un gráfico de control EWMA con un tamaño de los subgrupos igual a 4 y con un factor de ponderación $\beta = 0{,}25$. Si el valor actual del gráfico es 60 y las 16 siguientes medias de los subgrupos coinciden con

$$48, 52, 70, 62, 57, 81, 56, 59, 77, 82, 78, 80, 74, 82, 68, 84$$

¿qué se puede concluir?

Solución Si se comienza con $W_0 = 60$, los sucesivos valores de $W_1, \ldots, W_{16}$ se pueden calcular a partir de la expresión

$$W_t = 0{,}25\overline{X}_t + 0{,}75W_{t-1}$$

Así se obtiene que

$$\begin{aligned}
W_1 &= (0{,}25)(48) + (0{,}75)(60) = 57\\
W_2 &= (0{,}25)(52) + (0{,}75)(57) = 55{,}75\\
W_3 &= (0{,}25)(70) + (0{,}75)(55{,}75) = 59{,}31\\
W_4 &= (0{,}25)(62) + (0{,}75)(59{,}31) = 59{,}98\\
W_5 &= (0{,}25)(57) + (0{,}75)(59{,}98) = 59{,}24\\
W_6 &= (0{,}25)(81) + (0{,}75)(59{,}24) = 64{,}68
\end{aligned}$$

y, de igual forma, los valores desde W_7 hasta W_{16} coinciden con

$$62{,}50, 61{,}61, 65{,}48, 69{,}60, 71{,}70, 73{,}78, 73{,}83, 75{,}87, 73{,}90, 76{,}43$$

En la figura 15.6 se representan gráficamente los sucesivos valores de las medias móviles. Puesto que

$$3\sqrt{\frac{0{,}25}{1{,}75}}\,\frac{24}{\sqrt{4}} = 13{,}61$$

los límites de control del gráfico de control EWMA estándar con un factor de ponderación $\beta = 0{,}025$ son

$$\begin{aligned}
\text{LCL} &= 62 - 13{,}61 = 48{,}39\\
\text{UCL} &= 62 + 13{,}61 = 75{,}61
\end{aligned}$$

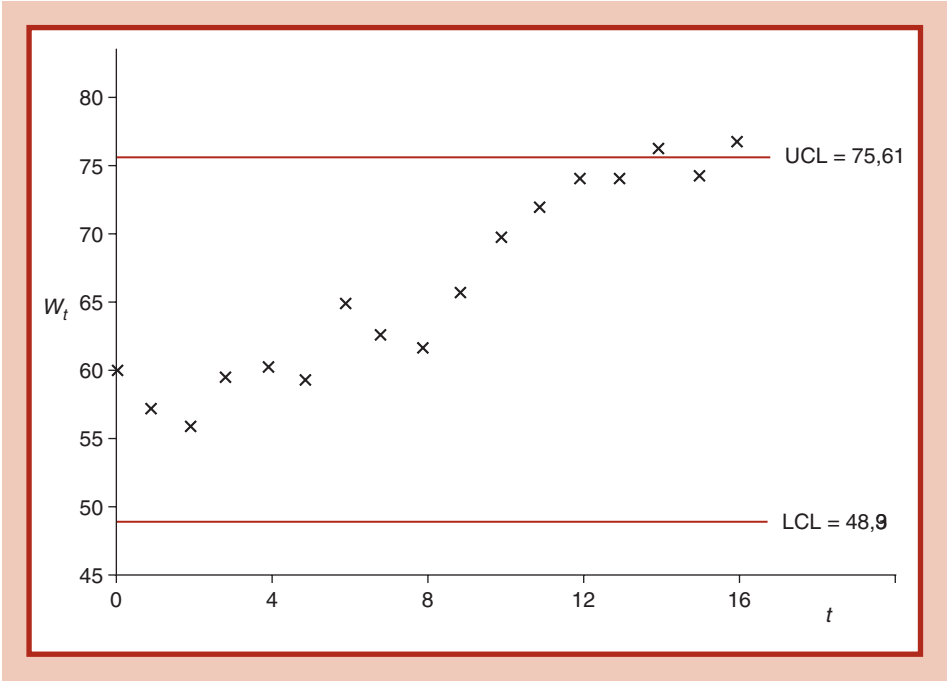


Figura 15.6 Un gráfico de control EWMA.

Así pues, el gráfico de control EWMA habría declarado que el sistema estuvo fuera de control tras obtenerse W_{14} (y también W_{16}). Por otra parte, dado que la desviación típica de un subgrupo es $\sigma/\sqrt{n} = 12$, es interesante observar que ningún valor de dato difiere de $\mu = 62$ más de 2 desviaciones típicas de los subgrupos, por cuya razón el gráfico de control de $\overline{X}$ no habría declarado que el sistema ha estado fuera de control. ■

Problemas

1. Consideremos un proceso cuyas medias de los subgrupos, cuando está bajo control, son normales con media 50 y desviación típica 5. La tabla siguiente representa 50 valores simulados procedentes de una distribución normal con media 54 y desviación típica 5. Esto es, los datos representan las medias de los subgrupos si el proceso está fuera de control debido a un incremento en la media de 0,8 desviaciones típicas de los subgrupos.

Subgrupos 1–10	Subgrupos 11–20	Subgrupos 21–30	Subgrupos 31–40	Subgrupos 41–50
59,81	52,79	50,04	58,56	54,95
51,38	56,52	51,49	50,79	53,22
52,87	54,98	56,93	54,19	46,21
62,38	58,00	50,13	58,65	47,08
53,59	56,91	52,29	57,41	53,31
60,39	54,26	56,74	56,79	51,42
46,64	58,43	48,62	50,86	61,57
55,84	53,41	44,58	51,14	57,33
55,78	48,43	57,46	61,26	60,34
53,27	52,65	60,98	56,68	49,97

(a) ¿El gráfico de control para $\overline{X}$ habría descubierto que el proceso se encuentra fuera de control?

(b) Repita el apartado (a), utilizando ahora un gráfico EWMA estándar con $\beta = 0{,}5$.

(c) Repita el apartado (a), utilizando ahora un gráfico EWMA estándar con $\beta = 0{,}25$.

2. Repita el problema 1 y en esta ocasión utilice los datos en sentido inverso. Esto es, tome 49,97 como primer valor, 60,34 como segundo, y así sucesivamente.

3. Repita el problema 2 de la sección 15.2, y en este caso emplee un gráfico de control EWMA estándar con un factor de ponderación $\beta = 0{,}7$.

4. Repita el problema 2 de la sección 15.2, y utilice un gráfico de control EWMA estándar con un factor de ponderación $\beta = 0{,}5$.

15.5 Gráficos de control de sumas acumuladas

El mayor competidor de de los gráficos de control tipo EWMA para detectar cambios pequeños o moderados en las medias es el llamado gráfico de control de *sumas acumuladas* (a menudo conocido simplemente como *cu-sum*, del inglés *cumulative sum*).

Supongamos como antes que $\overline{X}_1, \overline{X}_2, \ldots$ representan las medias sucesivas de los subgrupos de tamaño n y que, cuando el proceso se encuentra bajo control, dichas variables aleatorias tienen media μ y desviación típica $\sigma/\sqrt{n}$. Inicialmente, supongamos que sólo estamos interesados en determinar cuándo se ha producido un aumento en el valor medio. El gráfico de control (unilateral) para detectar un incremento en la media opera como sigue. Elijamos unas constantes positivas d y B, y definamos

$$Y_j = \overline{X}_j - \mu - \frac{d\sigma}{\sqrt{n}} \qquad j \geq 1$$

Observe que cuando el proceso está bajo y, por tanto, $E[\overline{X}_j] = \mu$, se verifica que

$$E[Y_j] = -\frac{d\sigma}{\sqrt{n}} < 0$$

Hagamos

$$S_0 = 0$$

$$S_{j+1} = \max\{S_j + Y_{j+1}, 0\}, \qquad j \geq 0$$

El gráfico de control cu-sum de parámetros d y B representa sucesivamente los valores S_j y declara que el valor medio ha aumentado en el primer j para el que

$$S_j > \frac{B\sigma}{\sqrt{n}}$$

Para entender el razonamiento en el que se basa este tipo de gráfico, supongamos que hubiéramos decidido representar las sumas sucesivas de las variables aleatorias Y_i observadas hasta el momento presente. Esto es, supongamos que hubiéramos decidido representar los sucesivos valores de P_j, siendo

$$P_j = \sum_{i=1}^{j} Y_i$$

que también se pueden escribir mediante

$$P_0 = 0$$

$$P_{j+1} = P_j + Y_{j+1}, \qquad j \geq 0$$

Si el sistema hubiera estado siempre bajo control, todas las Y_i tendrían un valor esperado negativo y, en consecuencia, puede esperarse que su suma sea negativa. De aquí se desprende que, si en alguna ocasión el valor de P_j se hace grande –digamos, por encima de $B\sigma/\sqrt{n}$– , esto proporcionaría una fuerte evidencia de que el proceso se ha salido de control (porque se ha producido un aumento en el valor medio de los ítems producidos). Sin embargo, la dificultad es que, si el sistema se sale de control después de un largo periodo de tiempo, es muy probable que el valor de P_j en ese momento sea fuertemente negativo (puesto que hasta entonces se habrán estado sumando variables aleatorias de media negativa) y, en consecuencia, se tardará tiempo en que su valor exceda de $B\sigma/\sqrt{n}$. Por este motivo, para conseguir que esta suma no se haga muy negativa mientras que el proceso se encuentra bajo control, los gráficos de control de sumas acumuladas reasignan el valor 0 en cuanto la suma toma un valor negativo. Esto es, las variables S_j coinciden con las sumas acumuladas Y_j, con la excepción de que en cualquier momento en que esta suma se hace negativa su valor se reajusta a 0.

Ejemplo 15.7 Supongamos que la media y la desviación típica de las medias de los subgrupos son, respectivamente, $\mu = 30$ y $\sigma/\sqrt{n} = 8$ y consideremos el gráfico de control cu-sum con $d = 0{,}5$ y $B = 5$. Si las primeras ocho medias de subgrupos son

$$29, 33, 35, 42, 36, 44, 43, 45$$

los sucesivos valores de $Y_j = \bar{X}_j - 30 - 4 = \bar{X}_j - 34$ serán

$$Y_1 = -5, \quad Y_2 = -1, \quad Y_3 = 1, \quad Y_4 = 8, \quad Y_5 = 2, \quad Y_6 = 10, \quad Y_7 = 9, \quad Y_8 = 11$$

Por consiguiente,

$$S_1 = \max\{-5, 0\} = 0$$

$$S_2 = \max\{-1, 0\} = 0$$

$$S_3 = \max\{1, 0\} = 1$$

$$S_4 = \max\{9, 0\} = 9$$

$$S_5 = \max\{11, 0\} = 11$$

$$S_6 = \max\{21, 0\} = 21$$

$$S_7 = \max\{30, 0\} = 30$$

$$S_8 = \max\{41, 0\} = 41$$

Puesto que el límite de control es

$$\frac{B\sigma}{\sqrt{n}} = 5(8) = 40$$

el gráfico de sumas acumuladas, tras haber observado la media del octavo subgrupo, declarará que la media ha aumentado. ■

Para poder detectar si se ha producido un cambio en la media, positivo o negativo, se suelen emplear simultáneamente dos gráficos unilaterales de sumas acumuladas. Observe en primer lugar que una disminución en $E[X_i]$ es equivalente a un aumento en $E[-X_i]$. De aquí se desprende que una disminución en el valor medio de los ítems producidos si se lleva a cabo un gráfico cu-sum unilateral y se cambian de signo las medias de los subgrupos. Esto es, una vez prefijados los valores d y B, no únicamente se representan gráficamente los valores de S_j como antes, sino que, además, se calculan

$$W_j = -\bar{X}_j - (-\mu) - \frac{d\sigma}{\sqrt{n}} = \mu - \bar{X}_j - \frac{d\sigma}{\sqrt{n}}$$

y del mismo modo se representan gráficamente los valores T_j, siendo

$$T_0 = 0$$
$$T_{j+1} = \max\{T_j + W_{j+1}, 0\}, \qquad j \geq 0$$

En cuanto que algún S_j o algún T_j sobrepase $B\sigma/\sqrt{n}$ se declarará que el proceso se encuentra fuera de control.

Resumiendo: Se indican a continuación los pasos que se deben dar para detectar un cambio en la media de los ítems producidos mediante un gráfico de control de sumas acumuladas. Elija unas constantes positivas d y B; utilice las medias de los sucesivos subgrupos para obtener los valores de S_j y T_j; declare que el proceso está fuera de control cuando alguno de ellos supere $B\sigma/\sqrt{n}$. Las tres elecciones usuales de los pares de valores d y B son: $d = 0{,}25$, $B = 8{,}00$; $d = 0{,}50$, $B = 4{,}77$; $d = 1$, $B = 2{,}49$. Con cada uno de ellos se consigue un gráfico de control que tiene aproximadamente la misma tasa de alarmas falsas que el gráfico de control de $\overline{X}$ que declara al proceso fuera de control cuando la media de un subgrupo difiera de μ por encima de $3\sigma/\sqrt{n}$. Grosso modo, como regla general, cuanto menor sea el cambio del que uno quiera protegerse, menor debería ser el valor de d elegido.

Problemas

1. Repita el problema 1 de la sección 15.5, y en esta ocasión utilice un gráfico de sumas acumuladas con:

   (a) $d = 0{,}25$ y $B = 8{,}00$

   (b) $d = 0{,}50$ y $B = 4{,}77$

2. Repita el problema 2 de la sección 15.2, y esta vez utilice un gráfico de sumas acumuladas con $d = 1$ y $B = 2{,}49$.

3. Repita el problema 3 de la sección 15.2, y emplee un gráfico de sumas acumuladas con $d = 0{,}50$ y $B = 4{,}77$.

Términos clave

Gráfico de control: Un procedimiento gráfico que permite detectar cuando un proceso de producción se encuentra fuera de control.

Resumen

Consideremos un proceso de producción de ítems, sobre cada uno de los cuales se puede medir una variable que tiene una media μ y una desviación típica σ cuando el proceso está bajo control. Para detectar cualquier tipo de cambios, los ítems se colocan en subgrupos de tamaño n y se representan gráficamente las medias $\overline{X}$ de los subgrupos. Siempre que la media de un subgrupo sea menor que el límite de control inferior

$$\text{LCL} = \mu - \frac{3\sigma}{\sqrt{n}}$$

o mayor que el límite de control superior

$$\text{UCL} = \mu + \frac{3\sigma}{\sqrt{n}}$$

el proceso se declara fuera de control.

En ocasiones, en lugar de tener medidas de una variable continua, cada item se clasifica bien como aceptable o como defectuoso. Denotemos por p la probabilidad de que un item sea defectuoso cuando el proceso se encuentra bajo control. Para dictaminar cuándo se ha perdido el control, los ítems producidos se colocan de nuevo en subgrupos de tamaño n. En cuanto ocurra que el número de defectos de un subgrupo cae fuera de los límites de control

$$\text{LCL} = np - 3\sqrt{np(1-p)} \qquad \text{y} \qquad \text{UCL} = np + 3\sqrt{np(1-p)}$$

el proceso se declara fuera de control.

Otros tipos de gráficos de control considerados son los gráficos de control *de medias móviles ponderados exponencialmente* y los *de sumas acumuladas*. Los primeros representan gráficamente las medias ponderadas de todas las medias observadas de los subgrupos hasta el instante presente, decreciendo exponencialmente los pesos de las medias de los subgrupos más antiguos. Los segundos representan gráficamente las sumas acumuladas de términos cuya media sea negativa si el proceso se encuentra bajo control, reajustándose esta suma a 0 cuando toma un valor negativo. El proceso se diagnostica como fuera de control si la suma acumulada sobrepasa un valor prefijado de antemano.

Problemas de repaso

1. La distancia entre dos clavijas adyacentes de un chip de memoria para adaptadores gráficos intensificados tiene, cuando el proceso de producción se encuentra bajo control, una media de 1,5 milímetros y una desviación típica de 0,001 milímetros. Calcule los límites de control superior e inferior de un gráfico de control de $\overline{X}$, con subgrupos de tamaño 4.

2. Con anterioridad a 1993, el número de hurtos cometidos anualmente en Estados Unidos por cada 100 000 habitantes seguía una distribución normal con media 1236 y desviación típica 120. Los siguientes datos muestran dichas tasas entre 1993 y 1001:

   1099,7, 1042,1, 987,0, 945,0, 918,8, 863,2, 770,4, 728,8, 740,8

   ¿Se puede concluir que la tasa de hurtos ha cambiado con respecto a su valor histórico? Utilice subgrupos de tamaño 3.

3. Cuando un proceso está funcionando correctamente, un 1,5% de los ítems producidos no se ajustan a las especificaciones. Si los ítems se agrupan en subgrupos de tamaño 300, calcule los límites de control superior e inferior de su gráfico de control.

4. Los siguientes datos muestran los enchufes defectuosos detectados en 12 muestras de tamaño 200:

   4, 7, 2, 5, 9, 5, 7, 10, 8, 3, 12, 9

   Supongamos que, cuando el proceso está bajo control, cada enchufe es defectuoso con probabilidad 0,03. ¿Parece que el proceso ha estado bajo control?

Apéndices

A

Un conjunto de datos

Estudiante	Peso	Colesterol	Presión	Sexo	Estudiante	Peso	Colesterol	Presión	Sexo
1	147	213	127	M	29	132	171	112	H
2	156	174	116	H	30	129	194	114	H
3	112	193	110	M	31	111	184	104	M
4	127	196	110	M	32	156	191	118	H
5	144	220	130	M	33	155	221	107	M
6	140	183	99	H	34	104	212	111	M
7	119	194	112	M	35	217	221	156	H
8	139	200	102	M	36	132	204	117	M
9	161	192	121	H	37	103	204	121	M
10	146	200	125	M	38	171	191	105	H
11	190	200	125	H	39	135	183	110	M
12	126	199	133	M	40	249	227	137	H
13	164	178	130	H	41	185	188	119	H
14	176	183	136	H	42	194	200	109	H
15	131	188	112	M	43	165	197	123	H
16	107	193	113	M	44	121	208	100	M
17	116	187	112	M	45	124	218	102	M
18	157	181	129	H	46	113	194	119	M
19	186	193	137	H	47	110	212	119	M
20	189	205	113	H	48	136	207	99	M
21	147	196	113	H	49	221	219	149	H
22	112	211	110	M	50	151	201	109	M
23	209	202	97	H	51	182	208	130	H
24	135	213	103	M	52	151	192	107	H
25	168	216	95	H	53	182	192	136	H
26	209	206	107	H	54	149	191	124	H
27	102	195	102	M	55	162	196	132	H
28	166	191	111	H	56	168	193	92	H

Estudiante	Peso	Colesterol	Presión	Sexo	Estudiante	Peso	Colesterol	Presión	Sexo
57	185	185	123	H	98	180	198	123	H
58	191	201	118	H	99	130	180	94	H
59	173	185	114	H	100	130	204	118	M
60	186	203	114	H	101	150	197	110	M
61	161	177	119	H	102	184	192	129	H
62	149	213	124	M	103	179	202	129	H
63	103	192	104	M	104	105	211	109	M
64	126	193	99	M	105	157	179	109	H
65	181	212	141	H	106	202	210	124	H
66	190	188	124	H	107	140	188	112	M
67	124	201	114	M	108	165	203	114	M
68	175	219	125	H	109	184	199	151	H
69	161	189	120	H	110	132	195	129	M
70	160	203	108	M	111	119	202	117	M
71	171	186	111	H	112	158	195	112	H
72	176	186	114	H	113	138	217	101	M
73	156	196	99	H	114	177	194	136	H
74	126	195	123	M	115	99	204	129	M
75	138	205	113	M	116	177	198	126	H
76	136	223	131	M	117	134	195	111	M
77	192	195	125	H	118	133	168	98	H
78	122	205	110	M	119	194	201	120	H
79	176	198	96	H	120	140	211	132	M
80	195	215	143	H	121	104	195	106	M
81	126	202	102	M	122	191	180	130	H
82	138	196	124	M	123	184	205	116	H
83	166	196	103	H	124	155	189	117	H
84	86	190	106	M	125	126	196	112	M
85	90	185	110	M	126	190	195	124	H
86	177	188	109	H	127	132	218	120	M
87	136	197	129	M	128	133	194	121	M
88	103	196	95	M	129	174	203	128	H
89	190	227	134	H	130	168	190	120	H
90	130	211	119	M	131	190	196	132	H
91	205	219	130	H	132	176	194	107	H
92	127	202	121	M	133	121	210	118	M
93	182	204	129	H	134	131	167	105	H
94	122	213	116	M	135	174	203	88	H
95	139	202	102	M	136	112	183	94	M
96	189	205	102	H	137	121	203	116	M
97	147	184	114	H	138	132	194	104	M

Estudiante	Peso	Colesterol	Presión	Sexo	Estudiante	Peso	Colesterol	Presión	Sexo
139	155	188	111	H	180	114	200	109	M
140	127	189	106	M	181	125	206	135	M
141	151	193	120	H	182	129	214	100	M
142	189	221	126	H	183	115	207	115	M
143	123	194	129	M	184	142	197	118	M
144	137	196	113	M	185	183	202	114	H
145	122	201	113	M	186	181	212	118	H
146	126	212	121	M	187	108	185	96	M
147	136	210	120	M	188	126	194	122	M
148	145	168	115	H	189	175	201	138	H
149	202	202	122	H	190	168	182	118	H
150	151	206	108	M	191	115	194	122	M
151	137	178	128	H	192	129	193	90	M
152	90	178	100	M	193	131	209	119	M
153	177	220	123	H	194	187	182	134	H
154	139	214	120	M	195	185	200	127	H
155	172	191	117	H	196	114	196	113	M
156	107	179	106	M	197	206	216	124	H
157	186	209	129	H	198	151	212	113	M
158	198	196	140	H	199	128	204	110	M
159	113	184	110	M	200	128	204	115	M
160	143	209	105	M	201	183	190	136	H
161	205	198	137	H	202	104	192	93	M
162	186	206	111	H	203	99	209	110	M
163	174	189	129	H	204	201	208	120	H
164	171	197	132	H	205	129	204	100	M
165	209	202	128	H	206	149	193	117	M
166	126	203	134	M	207	123	200	120	M
167	160	185	109	H	208	179	191	122	H
168	127	212	124	M	209	150	216	128	M
169	112	193	115	M	210	133	193	110	M
170	155	184	112	H	211	112	190	107	M
171	111	181	111	M	212	175	188	113	H
172	151	196	129	H	213	120	182	126	M
173	110	181	113	M	214	126	207	110	M
174	159	192	115	H	215	170	201	101	H
175	173	196	131	H	216	175	211	115	H
176	148	191	101	H	217	134	219	129	M
177	141	216	110	M	218	118	211	113	M
178	161	186	123	H	219	118	178	109	M
179	125	209	113	M	220	164	196	107	H

Estudiante	Peso	Colesterol	Presión	Sexo	Estudiante	Peso	Colesterol	Presión	Sexo
221	186	190	134	H	261	122	199	107	M
222	172	189	134	H	262	177	207	124	H
223	173	207	101	H	263	184	204	122	H
224	185	206	128	H	264	113	198	121	M
225	190	198	117	H	265	214	221	142	H
226	146	200	112	M	266	144	205	111	H
227	103	179	100	M	267	188	188	132	H
228	124	215	124	M	268	114	204	127	M
229	186	213	124	H	269	158	213	111	M
230	166	166	129	H	270	146	196	116	H
231	138	201	120	M	271	195	195	148	H
232	175	198	118	H	272	199	201	125	H
233	104	194	100	M	273	148	202	120	M
234	213	206	130	H	274	164	190	113	H
235	171	182	118	H	275	137	196	107	M
236	180	213	119	H	276	133	173	121	H
237	187	197	128	H	277	104	214	112	M
238	117	194	106	M	278	126	194	116	M
239	108	185	105	M	279	120	220	116	M
240	128	202	105	M	280	148	204	131	M
241	170	196	118	H	281	100	206	89	M
242	183	176	126	H	282	178	190	125	H
243	143	190	101	H	283	149	188	108	M
244	160	205	120	M	284	157	194	124	H
245	185	184	113	H	285	99	203	95	M
246	122	193	142	M	286	192	208	127	H
247	225	218	142	H	287	175	181	145	H
248	139	191	99	M	288	208	193	123	H
249	123	207	116	M	289	201	208	138	H
250	129	176	108	M	290	174	199	111	H
251	142	220	137	M	291	188	189	119	H
252	146	191	116	H	292	151	205	133	M
253	129	201	100	M	293	202	220	126	H
254	163	171	119	H	294	125	198	106	M
255	177	206	134	H	295	176	190	116	H
256	183	190	116	H	296	183	188	96	H
257	120	201	104	M	297	118	198	130	M
258	188	214	115	H	298	125	204	111	M
259	140	182	119	H	299	237	209	127	H
260	166	197	113	H	300	124	186	127	M

Estudiante	Peso	Colesterol	Presión	Sexo	Estudiante	Peso	Colesterol	Presión	Sexo
301	98	194	104	M	307	225	212	142	H
302	182	199	108	H	308	181	200	122	H
303	184	206	149	H	309	178	187	121	H
304	137	189	113	M	310	132	221	110	M
305	126	177	111	M	311	164	201	134	H
306	202	198	130	H	312	163	191	138	H

B

Preliminares matemáticos

B.1 Sumatorios

Consideremos cuatro números que se denotarán por x_1, x_2, x_3 y x_4. Si la suma de estos cuatro números es s, este hecho se expresará mediante

$$s = x_1 + x_2 + x_3 + x_4$$

o bien cuando el símbolo Σ. En este último caso, se escribirá

$$s = \sum_{i=1}^{4} x_i,$$

lo que significa que s es igual a la suma de los valores x_i, cuando i varía de 1 a 4.

El símbolo sumatorio suele ser útil si se desea sumar una gran cantidad de valores. Por ejemplo, supongamos que disponemos de 100 valores denotados por x_1, x_2, y así sucesivamente hasta x_{100}. Se puede representar la suma de estos 100 valores mediante

$$s = \sum_{i=1}^{100} x_i$$

Si se desea sumar únicamente los términos que empiezan en x_{20} y terminan en x_{79}, la suma resultante se representará de la forma

$$\sum_{i=20}^{79} x_i$$

Esto es, $\sum_{i=20}^{79} x_i$ denota la suma de los valores x_i, cuando i varía entre 20 y 79.

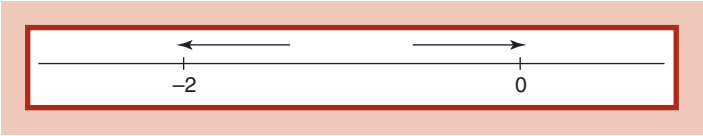


Figura B.1 La distancia entre -2 y 0 es $|-2| = 2$.

B.2 Valor absoluto

El valor absoluto de un número coincide con su magnitud sin considerar el signo. Por ejemplo, el valor absoluto de 4 es 4, mientras que el valor absoluto de -5 es 5. En general, el valor absoluto de un número positivo coincide con el número, mientras que el valor absoluto de un número negativo es su opuesto. Se utiliza el símbolo $|x|$ para denotar el valor absoluto del número x. Así pues,

$$|x| = \begin{cases} x & \text{si } x \geq 0 \\ -x & \text{si } x < 0 \end{cases}$$

Si cada número real se representa por un punto de la recta, $|x|$ representa la distancia desde el punto x al origen 0. Esto se ilustra en la figura B.1.

Si x e y son dos números, $|x - y|$ es igual a la distancia entre x e y. Por ejemplo, si $x = 5$ e $y = 2$, $|x - y| = |5 - 2| = |3| = 3$. Por otra parte, si $x = 5$ e $y = -2$, $|x - y| = |5 - (-2)| = |5 + 2| = 7$. Esto es, la distancia entre 5 y 2 es 3, mientras que la distancia entre 5 y -2 es 7.

B.3 Notación de conjuntos

Consideremos una colección de números, por ejemplo, todos los números reales. En ocasiones, podemos estar interesados en una subcolección de tales números, formada por aquellos que cumplen una determinada propiedad. Designemos por A la propiedad citada; por ejemplo, A podría ser la propiedad de ser positivo, o la de ser un entero par, o la de ser un entero primo. Se representarán los números de la subcolección definida por la propiedad A mediante la notación:

$$\{x\text{: } x \text{ cumple la propiedad } A\}$$

que se lee como "el conjunto de todos los valores x de la colección que cumplen la propiedad A." Por ejemplo,

$$\{x\text{: } x \text{ es un entero par comprendido entre 1 y 7}\}$$

es el conjunto formado por los tres valores 2, 4 y 6. Esto es,

$$\{x\text{: } x \text{ es un entero par comprendido entre 1 y 7}\} = \{2, 4, 6\}$$

En ocasiones, podemos estar interesados en el conjunto de todos los números que distan de un número dado menos de un determinado valor. Por ejemplo, consideremos el conjunto de todos los números que distan de 5 como máximo 2. Este conjunto se expresa como

$$\{x: |x - 5| \leq 2\}$$

Puesto que un número dista de 5 como máximo 2 si y sólo si dicho número está comprendido entre 3 y 7, se puede escribir

$$\{x: |x - 5| \leq 2\} = \{x: 3 \leq x \leq 7\}$$

C

Cómo seleccionar una muestra aleatoria

Como se ha visto a lo largo del libro es muy importante el poder seleccionar una muestra aleatoria. Supongamos que se desea extraer una muestra aleatoria de tamaño n procedente de una población de tamaño N. ¿Cómo se puede conseguir esto?

El primer paso consiste en numerar los miembros de la población de 1 a N de una forma cualquiera. Después se seleccionará la muestra aleatoria y se designarán los n elementos de la población que van a formar parte de la muestra. Para hacer esto se comienza permitiendo que el primer elemento de la muestra pueda ser uno cualquiera de los N elementos con la misma probabilidad. El segundo elemento de la muestra se elige de forma que los $N - 1$ elementos restantes de la población tengan la misma probabilidad de ser elegidos; el tercer elemento de la muestra puede ser uno cualquiera de los $N - 2$ elementos restantes, todos con la misma probabilidad; y así sucesivamente hasta que hayamos completado los n elementos que formarán parte de la muestra.

Para llevar a cabo este proceso podría parecer que siempre se tiene que seguir la pista de todos los individuos que han sido seleccionados previamente. Sin embargo, se puede ver muy sencillamente que esto no es necesario. Basta con colocar los N individuos de la población en una lista ordenada y luego se eligen no los individuos sino las posiciones que ocupan los individuos que van a formar parte de la muestra. Veamos cómo se hace cuando $N = 7$ y $n = 3$. Empezamos numerando cada uno de los 7 elementos de la población y luego los colocamos en una lista. Por ejemplo, supongamos que el orden inicial es

$$1, 2, 3, 4, 5, 6, 7$$

A continuación se selecciona un número de forma que 1, 2, 3, 4, 5, 6 y 7 tengan la misma probabilidad de ser elegidos; supongamos que 4 resulta elegido. Esto significa que el elemento que ocupa la posición 4 (el individuo 4 en este caso) formará parte de la muestra. Para indicar que este individuo está en la muestra aleatoria y para asegurarnos de que no volverá a resultar elegido, se intercambia el individuo que ocupa la posición 4 con el que ocupa la posición 7. Así se obtiene la nueva lista

$$1, 2, 3, 7, 5, 6, \underline{4,}$$

en la que se ha subrayado el elemento que ya está en la muestra. Para determinar cuál será el siguiente individuo muestral se seleccionará uno cualquiera de los que ocupan las 6 primeras posiciones de la lista nueva, siendo todos ellos igualmente probables. Así, se seleccionará un valor que puede coincidir con 1, 2, 3, 4, 5 o 6, todos con la misma probabilidad; el elemento que ocupe la posición resultante será el que formará parte de la muestra. Para indicar esto y para dejar que las primeras 5 posiciones queden ocupadas por los elementos aún no seleccionados se intercambian el elemento de la posición seleccionada y el de la sexta. Por ejemplo, si el valor elegido fue el 4, el elemento que ocupa la posición 4 (es decir, el elemento número 7) formará parte de la muestra, y la nueva lista será

$$1, 2, 3, 6, 5, 7, 4$$

Para extraer el último miembro de la muestra, los individuos que ocupan las cinco primeras posiciones de la lista son igualmente probables; así pues, se seleccionara equiprobablemente una cualquiera de las posiciones 1, 2, 3, 4 y 5, tras lo que se intercambian los elementos que ocupan la posición elegida y la quinta. Por ejemplo, si la posición 2 resultó elegida, la nueva lista será

$$1, 5, 3, 6, 2, 7, 4$$

Puesto que ya se han seleccionado los tres miembros de la muestra (2, 7 y 4), el proceso ha terminado.

Para implementar el anterior *algoritmo* de selección muestral se necesita saber cómo se puede generar un valor de una variable aleatoria que puede tomar los valores $1, 2, 3, \ldots, k$ con probabilidades iguales. El punto clave para hacerlo consiste en utilizar los *números aleatorios*, que son los valores de las variables aleatorias distribuidas uniformemente en el intervalo (0, 1). La mayor parte de los ordenadores disponen de un generador interno de números aleatorios con el que se puede obtener un valor de dichas variables. Si un número aleatorio se representa como U —esto es, si U sigue una distribución uniforme en el intervalo (0, 1)—, es fácil comprobar que

$$I = \text{Ent}\,(kU) + 1,$$

puede tomar los valores $1, 2, \ldots, k$ con la misma probabilidad, donde Ent (x) denota la parte entera de x. Por ejemplo,

$$\text{Ent}\,(4{,}3) = 4$$
$$\text{Ent}\,(12{,}9) = 12$$

y así sucesivamente.

El Programa A-1 utiliza todo lo anterior para extraer una muestra aleatoria de tamaño n del conjunto de números $1, 2, \ldots, N$. Para ejecutarlo, el programa nos pide que introduzcamos, en primer lugar, los valores de n y N, y, posteriormente, un número de cuatro dígitos. Para esto, basta con teclear por pantalla el primer número que nos venga a la mente. La salida del programa proporciona el conjunto de tamaño n que constituye la muestra aleatoria.

Ejemplo C.1 Supongamos que se desea seleccionar una muestra aleatoria de tamaño 12 procedente de una población con 200 miembros. Para hacerlo se empieza numerando los 200 elementos de la población de forma que todos ellos queden identificados mediante los números comprendidos entre 1 y 200. Con el Programa A-1, para extraer los 12 individuos de la población que formarán parte de la muestra, se obtiene lo siguiente:

```
THIS PROGRAM GENERATES A RANDOM SAMPLE OF K
OF THE INTEGERS 1 THRU N
ENTER THE VALUE OF N
? 200
ENTER THE VALUE OF K
? 12
Random Number Seed (−32,768 to 32,767)? 355
THE RANDOM SAMPLE CONSISTS OF THE FOLLOWING
12 ELEMENTS
90 89 82 162 21 81 182 45 38 195 64 1
```

■

D

Tablas

Tabla D.1 Probabilidades de la normal estándar

Los valores de la tabla representan $P\{Z \leq x\}$.

x	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0, 8186	0, 8212	0, 8238	0, 8264	0, 8289	0, 8315	0, 8340	0, 8365	0, 8389
1, 0	0, 8413	0, 8438	0, 8461	0, 8485	0, 8508	0, 8531	0, 8554	0, 8577	0, 8599	0, 8621
1, 1	0, 8643	0, 8665	0, 8686	0, 8708	0, 8729	0, 8749	0, 8770	0, 8790	0, 8810	0, 8830
1, 2	0, 8849	0, 8869	0, 8888	0, 8907	0, 8925	0, 8944	0, 8962	0, 8980	0, 8997	0, 9015
1, 3	0, 9032	0, 9049	0, 9066	0, 9082	0, 9099	0, 9115	0, 9131	0, 9147	0, 9162	0, 9177
1, 4	0, 9192	0, 9207	0, 9222	0, 9236	0, 9251	0, 9265	0, 9279	0, 9292	0, 9306	0, 9319
1, 5	0, 9332	0, 9345	0, 9357	0, 9370	0, 9382	0, 9394	0, 9406	0, 9418	0, 9429	0, 9441
1, 6	0, 9452	0, 9463	0, 9474	0, 9484	0, 9495	0, 9505	0, 9515	0, 9525	0, 9535	0, 9545
1, 7	0, 9554	0, 9564	0, 9573	0, 9582	0, 9591	0, 9599	0, 9608	0, 9616	0, 9625	0, 9633
1, 8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916

Tabla D.1 Probabilidades de la normal estándar (*continuación*)

x	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998

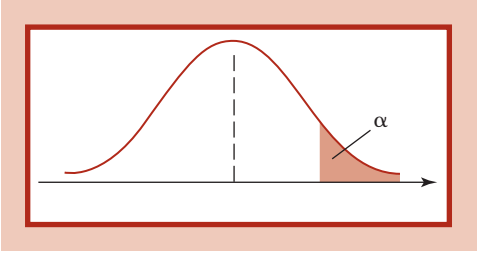


Tabla D.2 Percentiles $t_{n,\alpha}$ de las distribuciones t

	α									
n	**0,40**	**0,25**	**0,10**	**0,05**	**0,025**	**0,01**	**0,005**	**0,0025**	**0,001**	**0,0005**
1	0,325	1,000	3,078	6,314	12,706	31,821	63,657	127,32	318,31	636,62
2	0,289	0,816	1,886	2,920	4,303	6,965	9,925	14,089	23,326	31,598
3	0,277	0,765	1,638	2,353	3,182	4,541	5,841	7,453	10,213	12,924
4	0,271	0,741	1,533	2,132	2,776	3,747	4,604	5,598	7,173	8,610
5	0,267	0,727	1,476	2,015	2,571	3,365	4,032	4,773	5,893	6,869
6	0,265	0,718	1,440	1,943	2,447	3,143	3,707	4,317	5,208	5,959
7	0,263	0,711	1,415	1,895	2,365	2,998	3,499	4,029	4,785	5,408
8	0,262	0,706	1,397	1,860	2,306	2,896	3,355	3,833	4,501	5,041
9	0,261	0,703	1,383	1,833	2,262	2,821	3,250	3,690	4,297	4,781
10	0,260	0,700	1,372	1,812	2,228	2,764	3,169	3,581	4,144	4,587
11	0,260	0,697	1,363	1,796	2,201	2,718	3,106	3,497	4,025	4,437
12	0,259	0,695	1,356	1,782	2,179	2,681	3,055	3,428	3,930	4,318
13	0,259	0,694	1,350	1,771	2,160	2,650	3,012	3,372	3,852	4,221

Tabla D.2 Percentiles $t_{n,\alpha}$ de las distribuciones t (*continuación*)

	α									
n	**0,40**	**0,25**	**0,10**	**0,05**	**0,025**	**0,01**	**0,005**	**0,0025**	**0,001**	**0,0005**
14	0,258	0,692	1,345	1,761	2,145	2,624	2,977	3,326	3,787	4,140
15	0,258	0,691	1,341	1,753	2,131	2,602	2,947	3,286	3,733	4,073
16	0,258	0,690	1,337	1,746	2,120	2,583	2,921	3,252	3,686	4,015
17	0,257	0,689	1,333	1,740	2,110	2,567	2,898	3,222	3,646	3,965
18	0,257	0,688	1,330	1,734	2,101	2,552	2,878	3,197	3,610	3,922
19	0,257	0,688	1,328	1,729	2,093	2,539	2,861	3,174	3,579	3,883
20	0,257	0,687	1,325	1,725	2,086	2,528	2,845	3,153	3,552	3,850
21	0,257	0,686	1,323	1,721	2,080	2,518	2,831	3,135	3,527	3,819
22	0,256	0,686	1,321	1,717	2,074	2,508	2,819	3,119	3,505	3,792
23	0,256	0,685	1,319	1,714	2,069	2,500	2,807	3,104	3,485	3,767
24	0,256	0,685	1,318	1,711	2,064	2,492	2,797	3,091	3,467	3,745
25	0,256	0,684	1,316	1,708	2,060	2,485	2,787	3,078	3,450	3,725
26	0,256	0,684	1,315	1,706	2,056	2,479	2,779	3,067	3,435	3,707
27	0,256	0,684	1,314	1,703	2,052	2,473	2,771	3,057	3,421	3,690
28	0,256	0,683	1,313	1,701	2,048	2,467	2,763	3,047	3,408	3,674
29	0,256	0,683	1,311	1,699	2,045	2,462	2,756	3,038	3,396	3,659
30	0,256	0,683	1,310	1,697	2,042	2,457	2,750	3,030	3,385	3,646
40	0,255	0,681	1,303	1,684	2,021	2,423	2,704	2,971	3,307	3,551
60	0,254	0,679	1,296	1,671	2,000	2,390	2,660	2,915	3,232	3,460
120	0,254	0,677	1,289	1,658	1,980	2,358	2,617	2,860	3,160	3,373
∞	0,253	0,674	1,282	1,645	1,960	2,326	2,576	2,807	3,090	3,291

n = grados de libertad.

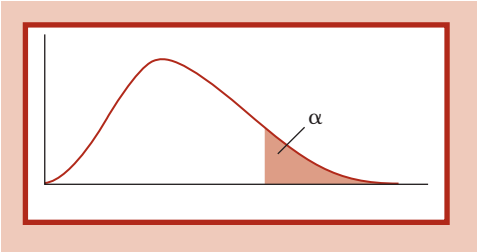


Tabla D.3 Percentiles $\chi^2_{n,\alpha}$ de las distribuciones chi-cuadrado

	α										
n	**0,995**	**0,990**	**0,975**	**0,950**	**0,900**	**0,500**	**0,100**	**0,050**	**0,025**	**0,010**	**0,005**
1	0,00+	0,00+	0,00+	0,00+	0,02	0,45	2,71	3,84	5,02	6,63	7,88
2	0,01	0,02	0,05	0,10	0,21	1,39	4,61	5,99	7,38	9,21	10,60
3	0,07	0,11	0,22	0,35	0,58	2,37	6,25	7,81	9,35	11,34	12,84
4	0,21	0,30	0,48	0,71	1,06	3,36	7,78	9,49	11,14	13,28	14,86
5	0,41	0,55	0,83	1,15	1,61	4,35	9,24	11,07	12,83	15,09	16,75
6	0,68	0,87	1,24	1,64	2,20	5,35	10,65	12,59	14,45	16,81	18,55
7	0,99	1,24	1,69	2,17	2,83	6,35	12,02	14,07	16,01	18,48	20,28
8	1,34	1,65	2,18	2,73	3,49	7,34	13,36	15,51	17,53	20,09	21,96
9	1,73	2,09	2,70	3,33	4,17	8,34	14,68	16,92	19,02	21,67	23,59
10	2,16	2,56	3,25	3,94	4,87	9,34	15,99	18,31	20,48	23,21	25,19
11	2,60	3,05	3,82	4,57	5,58	10,34	17,28	19,68	21,92	24,72	26,76
12	3,07	3,57	4,40	5,23	6,30	11,34	18,55	21,03	23,34	26,22	28,30
13	3,57	4,11	5,01	5,89	7,04	12,34	19,81	22,36	24,74	27,69	29,82
14	4,07	4,66	5,63	6,57	7,79	13,34	21,06	23,68	26,12	29,14	31,32
15	4,60	5,23	6,27	7,26	8,55	14,34	22,31	25,00	27,49	30,58	32,80
16	5,14	5,81	6,91	7,96	9,31	15,34	23,54	26,30	28,85	32,00	34,27
17	5,70	6,41	7,56	8,67	10,09	16,34	24,77	27,59	30,19	33,41	35,72
18	6,26	7,01	8,23	9,39	10,87	17,34	25,99	28,87	31,53	34,81	37,16
19	6,84	7,63	8,91	10,12	11,65	18,34	27,20	30,14	32,85	36,19	38,58
20	7,43	8,26	9,59	10,85	12,44	19,34	28,41	31,41	34,17	37,57	40,00
21	8,03	8,90	10,28	11,59	13,24	20,34	29,62	32,67	35,48	38,93	41,40
22	8,64	9,54	10,98	12,34	14,04	21,34	30,81	33,92	36,78	40,29	42,80
23	9,26	10,20	11,69	13,09	14,85	22,34	32,01	35,17	38,08	41,64	44,18
24	9,89	10,86	12,40	13,85	15,66	23,34	33,20	36,42	39,36	42,98	45,56
25	10,52	11,52	13,12	14,61	16,47	24,34	34,28	37,65	40,65	44,31	46,93
26	11,16	12,20	13,84	15,38	17,29	25,34	35,56	38,89	41,92	45,64	48,29
27	11,81	12,88	14,57	16,15	18,11	26,34	36,74	40,11	43,19	46,96	49,65
28	12,46	13,57	15,31	16,93	18,94	27,34	37,92	41,34	44,46	48,28	50,99
29	13,12	14,26	16,05	17,71	19,77	28,34	39,09	42,56	45,72	49,59	52,34
30	13,79	14,95	16,79	18,49	20,60	29,34	40,26	43,77	46,98	50,89	53,67

Tabla D.3 Percentiles $\chi^2_{n,\alpha}$ de las distribuciones chi-cuadrado (*continuación*)

	α										
n	**0,995**	**0,990**	**0,975**	**0,950**	**0,900**	**0,500**	**0,100**	**0,050**	**0,025**	**0,010**	**0,005**
40	20,71	22,16	24,43	26,51	29,05	39,34	51,81	55,76	59,34	63,69	66,77
50	27,99	29,71	32,36	34,76	37,69	49,33	63,17	67,50	71,42	76,15	79,49
60	35,53	37,48	40,48	43,19	46,46	59,33	74,40	79,08	83,30	88,38	91,95
70	43,28	45,44	48,76	51,74	55,33	69,33	85,53	90,53	95,02	100,42	104,22
80	51,17	53,54	57,15	60,39	64,28	79,33	96,58	101,88	106,63	112,33	116,32
90	59,20	61,75	65,65	69,13	73,29	89,33	107,57	113,14	118,14	124,12	128,30
100	67,33	70,06	74,22	77,93	82,36	99,33	118,50	124,34	129,56	135,81	140,17

n = grados de libertad

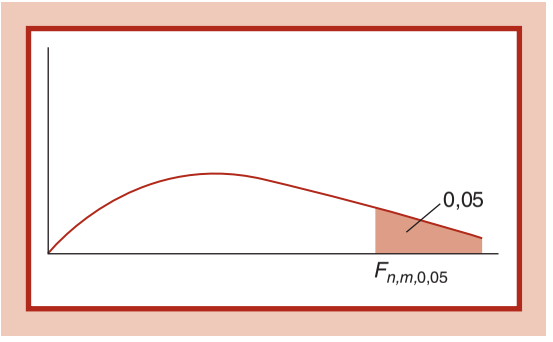


Tabla D.4 Percentiles de las distribuciones F

Percentiles del 95% de las distribuciones $F_{n,\,m}$

Grados de libertad del denominador m	Grados de libertad del numerador n																		
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	∞
1	161,4	199,5	215,7	224,6	230,2	234,0	236,8	238,9	240,5	241,9	243,9	245,9	248,0	249,1	250,1	251,1	252,2	253,3	254,3
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07

Tabla D.4 Percentiles de las distribuciones *F* (*continuación*)

Percentiles del 95% de las distribuciones $F_{n,m}$

Grados de libertad del denominador *m*	Grados de libertad del numerador *n*																		
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	∞
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96	1,91	1,83	1,75	1,66	1,61	1,55	1,55	1,43	1,35	1,25
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00

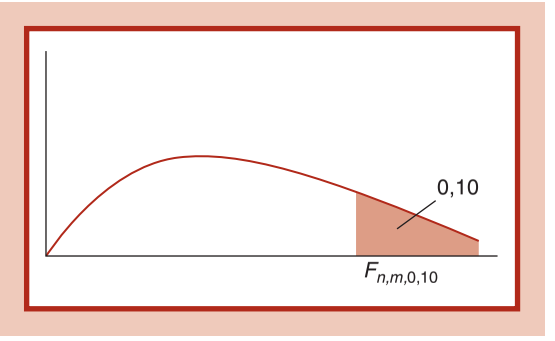


Tabla D.4 Percentiles de las distribuciones F (*continuación*)

Percentiles del 95% de las distribuciones $F_{n,\,m}$

Grados de libertad del denominador m	Grados de libertad del numerador n																		
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	∞
1	39,86	49,50	53,59	55,83	57,24	58,20	58,91	59,44	59,86	60,19	60,71	61,22	61,74	62,00	62,26	62,53	62,79	63,06	63,33
2	8,53	9,00	9,16	9,24	9,29	9,33	9,35	9,37	9,38	9,39	9,41	9,42	9,44	9,45	9,46	9,47	9,47	9,48	9,49
3	5,54	5,46	5,39	5,34	5,31	5,28	5,27	5,25	5,24	5,23	5,22	5,20	5,18	5,18	5,17	5,16	5,15	5,14	5,13
4	4,54	4,32	4,19	4,11	4,05	4,01	3,98	3,95	3,94	3,92	3,90	3,87	3,84	3,83	3,82	3,80	3,79	3,78	3,76
5	4,06	3,78	3,62	3,52	3,45	3,40	3,37	3,34	3,32	3,30	3,27	3,24	3,21	3,19	3,17	3,16	3,14	3,12	3,10
6	3,78	3,46	3,29	3,18	3,11	3,05	3,01	2,98	2,96	2,94	2,90	2,87	2,84	2,82	2,80	2,78	2,76	2,74	2,72
7	3,59	3,26	3,07	2,96	2,88	2,83	2,78	2,75	2,72	2,70	2,67	2,63	2,59	2,58	2,56	2,54	2,51	2,49	2,47
8	3,46	3,11	2,92	2,81	2,73	2,67	2,62	2,59	2,56	2,54	2,50	2,46	2,42	2,40	2,38	2,36	2,34	2,32	2,29
9	3,36	3,01	2,81	2,69	2,61	2,55	2,51	2,47	2,44	2,42	2,38	2,34	2,30	2,28	2,25	2,23	2,21	2,18	2,16
10	3,29	2,92	2,73	2,61	2,52	2,46	2,41	2,38	2,35	2,32	2,28	2,24	2,20	2,18	2,16	2,13	2,11	2,08	2,06
11	3,23	2,86	2,66	2,54	2,45	2,39	2,34	2,30	2,27	2,25	2,21	2,17	2,12	2,10	2,08	2,05	2,03	2,00	1,97
12	3,18	2,81	2,61	2,48	2,39	2,33	2,28	2,24	2,21	2,19	2,15	2,10	2,06	2,04	2,01	1,99	1,96	1,93	1,90
13	3,14	2,76	2,56	2,43	2,35	2,28	2,23	2,20	2,16	2,14	2,10	2,05	2,01	1,98	1,96	1,93	1,90	1,88	1,85
14	3,10	2,73	2,52	2,39	2,31	2,24	2,19	2,15	2,12	2,10	2,05	2,01	1,96	1,94	1,91	1,89	1,86	1,83	1,80
15	3,07	2,70	2,49	2,36	2,27	2,21	2,16	2,12	2,09	2,06	2,02	1,97	1,92	1,90	1,87	1,85	1,82	1,79	1,76

Tabla D.4 Percentiles de las distribuciones *F* (*continuación*)

Percentiles del 95% de las distribuciones $F_{n,m}$

Grados de libertad del denominador *m*	Grados de libertad del numerador *n*																		
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	∞
16	3,05	2,67	2,46	2,33	2,24	2,18	2,13	2,09	2,06	2,03	1,99	1,94	1,89	1,87	1,84	1,81	1,78	1,75	1,72
17	3,03	2,64	2,44	2,31	2,22	2,15	2,10	2,06	2,03	2,00	1,96	1,91	1,86	1,84	1,81	1,78	1,75	1,72	1,69
18	3,01	2,62	2,42	2,29	2,20	2,13	2,08	2,04	2,00	1,98	1,93	1,89	1,84	1,81	1,78	1,75	1,72	1,69	1,66
19	2,99	2,61	2,40	2,27	2,18	2,11	2,06	2,02	1,98	1,96	1,91	1,86	1,81	1,79	1,76	1,73	1,70	1,67	1,63
20	2,97	2,59	2,38	2,25	2,16	2,09	2,04	2,00	1,96	1,94	1,89	1,84	1,79	1,77	1,74	1,71	1,68	1,64	1,61
21	2,96	2,57	2,36	2,23	2,14	2,08	2,02	1,98	1,95	1,92	1,87	1,83	1,78	1,75	1,72	1,69	1,66	1,62	1,59
22	2,95	2,56	2,35	2,22	2,13	2,06	2,01	1,97	1,93	1,90	1,86	1,81	1,76	1,73	1,70	1,67	1,64	1,60	1,57
23	2,94	2,55	2,34	2,21	2,11	2,05	1,99	1,95	1,92	1,89	1,84	1,80	1,74	1,72	1,69	1,66	1,62	1,59	1,55
24	2,93	2,54	2,33	2,19	2,10	2,04	1,98	1,94	1,91	1,88	1,83	1,78	1,73	1,70	1,67	1,64	1,61	1,57	1,53
25	2,92	2,53	2,32	2,18	2,09	2,02	1,97	1,93	1,89	1,87	1,82	1,77	1,72	1,69	1,66	1,63	1,59	1,56	1,52
26	2,91	2,52	2,31	2,17	2,08	2,01	1,96	1,92	1,88	1,86	1,81	1,76	1,71	1,68	1,65	1,61	1,58	1,54	1,50
27	2,90	2,51	2,30	2,17	2,07	2,00	1,95	1,91	1,87	1,85	1,80	1,75	1,70	1,67	1,64	1,60	1,57	1,53	1,49
28	2,89	2,50	2,29	2,16	2,06	2,00	1,94	1,90	1,87	1,84	1,79	1,74	1,69	1,66	1,63	1,59	1,56	1,52	1,48
29	2,89	2,50	2,28	2,15	2,06	1,99	1,93	1,89	1,86	1,83	1,78	1,73	1,68	1,65	1,62	1,58	1,55	1,51	1,47
30	2,88	2,49	2,28	2,14	2,03	1,98	1,93	1,88	1,85	1,82	1,77	1,72	1,67	1,64	1,61	1,57	1,54	1,50	1,46
40	2,84	2,44	2,23	2,09	2,00	1,93	1,87	1,83	1,79	1,76	1,71	1,66	1,61	1,57	1,54	1,51	1,47	1,42	1,38
60	2,79	2,39	2,18	2,04	1,95	1,87	1,82	1,77	1,74	1,71	1,66	1,60	1,54	1,51	1,48	1,44	1,40	1,35	1,29
120	2,75	2,35	2,13	1,99	1,90	1,82	1,77	1,72	1,68	1,65	1,60	1,55	1,48	1,45	1,41	1,37	1,32	1,26	1,19
∞	2,71	2,30	2,08	1,94	1,85	1,77	1,72	1,67	1,63	1,60	1,55	1,49	1,42	1,38	1,34	1,30	1,24	1,17	1,00

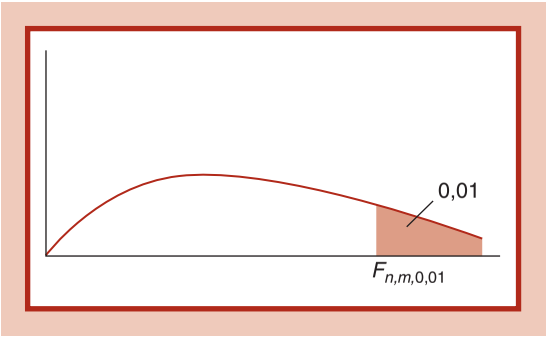


Tabla D.4 Percentiles de las distribuciones F (*continuación*)

Percentiles del 95% de las distribuciones $F_{n,\,m}$

Grados de libertad del denominador m	Grados de libertad del numerador n																		
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	∞
1	4052	4999,5	5403	5625	5764	5859	5928	5982	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,39	99,40	99,42	99,43	99,45	99,46	99,47	99,47	99,48	99,49	99,50
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23	27,05	26,87	26,69	26,00	26,50	26,41	26,32	26,22	26,13
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,37	14,20	14,02	13,93	13,84	13,75	13,65	13,56	13,46
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,89	9,72	9,55	9,47	9,38	9,29	9,20	9,11	9,02
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,72	7,56	7,40	7,31	7,23	7,14	7,06	6,97	6,88
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,47	6,31	6,16	6,07	5,99	5,91	5,82	5,74	5,65
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,67	5,52	5,36	5,28	5,20	5,12	5,03	4,95	4,46
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,11	4,96	4,81	4,73	4,65	4,57	4,48	4,40	4,31
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,71	4,56	4,41	4,33	4,25	4,17	4,08	4,00	3,91
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,40	4,25	4,10	4,02	3,94	3,86	3,78	3,69	3,60
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,16	4,01	3,86	3,78	3,70	3,62	3,54	3,45	3,36
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	3,96	3,82	3,66	3,59	3,51	3,43	3,34	3,25	3,17
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,80	3,66	3,51	3,43	3,35	3,27	3,18	3,09	3,00
15	8,68	6,36	5,42	4,89	4,36	4,32	4,14	4,00	3,89	3,80	3,67	3,52	3,37	3,29	3,21	3,13	3,05	2,96	2,87

Tabla D.4 Percentiles de las distribuciones *F* (*continuación*)

Percentiles del 95% de las distribuciones $F_{n,m}$

Grados de libertad del denominador *m*	Grados de libertad del numerador *n*																		
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	∞
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,55	3,41	3,26	3,18	3,10	3,02	2,93	2,84	2,75
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,46	3,31	3,16	3,08	3,00	2,92	2,83	2,75	2,65
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,37	3,23	3,08	3,00	2,92	2,84	2,75	2,66	2,57
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,30	3,15	3,00	2,92	2,84	2,76	2,67	2,58	2,59
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,23	3,09	2,94	2,86	2,78	2,69	2,61	2,52	2,42
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31	3,17	3,03	2,88	2,80	2,72	2,64	2,55	2,46	2,36
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,12	2,98	2,83	2,75	2,67	2,58	2,50	2,40	2,31
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21	3,07	2,93	2,78	2,70	2,62	2,54	2,45	2,35	2,26
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17	3,03	2,89	2,74	2,66	2,58	2,49	2,40	2,31	2,21
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13	2,99	2,85	2,70	2,62	2,54	2,45	2,36	2,27	2,17
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09	2,96	2,81	2,66	2,58	2,50	2,42	2,33	2,23	2,13
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06	2,93	2,78	2,63	2,55	2,47	2,38	2,29	2,20	2,10
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03	2,90	2,75	2,60	2,52	2,44	2,35	2,26	2,17	2,06
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00	2,87	2,73	2,57	2,49	2,41	2,33	2,23	2,14	2,03
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,11	2,01
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,66	2,52	2,37	2,29	2,20	2,11	2,02	1,92	1,80
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,50	2,35	2,20	2,12	2,03	1,94	1,84	1,73	1,60
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47	2,34	2,19	2,03	1,95	1,86	1,76	1,66	1,53	1,38
∞	6,63	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	2,18	2,04	1,88	1,79	1,70	1,59	1,47	1,32	1,00

Tabla D.5 Funciones de distribución binomiales

Los datos de la tabla reflejan los valores $P\{\text{Bin}(n, p) \leq i\}$, *donde* $\text{Bin}(n, p)$ *representa la variable aleatoria binomial de parámetros n y p. Para valores de* $p > 0{,}5$, *utilice la identidad* $P\{\text{Bin}(n, p) \leq i\} = 1 - P\{\text{Bin}(n, 1 - p) \leq n - i - 1\}$.

		p									
n	i	**0,05**	**0,10**	**0,15**	**0,20**	**0,25**	**0,30**	**0,35**	**0,40**	**0,45**	**0,50**
2	0	0,9025	0,8100	0,7225	0,6400	0,5625	0,4900	0,4225	0,3600	0,3025	0,2500
	1	0,9975	0,9900	0,9775	0,9600	0,9375	0,9100	0,8755	0,8400	0,7975	0,7500
3	0	0,8574	0,7290	0,6141	0,5120	0,4219	0,3430	0,2746	0,2160	0,1664	0,1250
	1	0,9928	0,9720	0,9392	0,8960	0,8438	0,7840	0,7182	0,6480	0,5748	0,5000
	2	0,9999	0,9990	0,9966	0,9920	0,9844	0,9730	0,9571	0,9360	0,9089	0,8750
4	0	0,8145	0,6561	0,5220	0,4096	0,3164	0,2401	0,1785	0,1296	0,0915	0,0625
	1	0,9860	0,9477	0,8905	0,8192	0,7383	0,6517	0,5630	0,4752	0,3910	0,3125
	2	0,9995	0,9963	0,9880	0,9728	0,9492	0,9163	0,8735	0,8208	0,7585	0,6875
	3	1,0000	0,9999	0,9995	0,9984	0,9961	0,9919	0,9850	0,9744	0,9590	0,9375
5	0	0,7738	0,5905	0,4437	0,3277	0,2373	0,1681	0,1160	0,0778	0,0503	0,0312
	1	0,9774	0,9185	0,8352	0,7373	0,6328	0,5282	0,4284	0,3370	0,2562	0,1875
	2	0,9988	0,9914	0,9734	0,9421	0,8965	0,8369	0,7648	0,6826	0,5931	0,5000
	3	1,0000	0,9995	0,9978	0,9933	0,9844	0,9692	0,9460	0,9130	0,8688	0,8125
	4	1,0000	1,0000	0,9999	0,9997	0,9990	0,9976	0,9947	0,9898	0,9815	0,9688
6	0	0,7351	0,5314	0,3771	0,2621	0,1780	0,1176	0,0754	0,0467	0,0277	0,0156
	1	0,9672	0,8857	0,7765	0,6554	0,5339	0,4202	0,3191	0,2333	0,1636	0,1094
	2	0,9978	0,9842	0,9527	0,9011	0,8306	0,7443	0,6471	0,5443	0,4415	0,3438
	3	0,9999	0,9987	0,9941	0,9830	0,9624	0,9295	0,8826	0,8208	0,7447	0,6562
	4	1,0000	0,9999	0,9996	0,9984	0,9954	0,9891	0,9777	0,9590	0,9308	0,8906
	5	1,0000	1,0000	1,0000	0,9999	0,9998	0,9993	0,9982	0,9959	0,9917	0,9844
7	0	0,6983	0,4783	0,3206	0,2097	0,1335	0,0824	0,0490	0,0280	0,0152	0,0078
	1	0,9556	0,8503	0,7166	0,5767	0,4449	0,3294	0,2338	0,1586	0,1024	0,0625
	2	0,9962	0,9743	0,9262	0,8520	0,7564	0,6471	0,5323	0,4199	0,3164	0,2266
	3	0,9998	0,9973	0,9879	0,9667	0,9294	0,8740	0,8002	0,7102	0,6083	0,5000
	4	1,0000	0,9998	0,9988	0,9953	0,9871	0,9712	0,9444	0,9037	0,8471	0,7734
	5	1,0000	1,0000	0,9999	0,9996	0,9987	0,9962	0,9910	0,9812	0,9643	0,9375
	6	1,0000	1,0000	1,0000	1,0000	0,9999	0,9998	0,9994	0,9984	0,9963	0,9922
8	0	0,6634	0,4305	0,2725	0,1678	0,1001	0,0576	0,0319	0,0168	0,0084	0,0039
	1	0,9428	0,8131	0,6572	0,5033	0,3671	0,2553	0,1691	0,1064	0,0632	0,0352
	2	0,9942	0,9619	0,8948	0,7969	0,6785	0,5518	0,4278	0,3154	0,2201	0,1445
	3	0,9996	0,9950	0,9786	0,9437	0,8862	0,8059	0,7064	0,5941	0,4770	0,3633
	4	1,0000	0,9996	0,9971	0,9896	0,9727	0,9420	0,8939	0,8263	0,7396	0,6367

Tabla D.5 Funciones de distribución binomiales (*continuación*)

		p									
n	***i***	**0,05**	**0,10**	**0,15**	**0,20**	**0,25**	**0,30**	**0,35**	**0,40**	**0,45**	**0,50**
	5	1,0000	1,0000	0,9998	0,9988	0,9958	0,9887	0,9747	0,9502	0,9115	0,8555
	6	1,0000	1,0000	1,0000	0,9999	0,9996	0,9987	0,9964	0,9915	0,9819	0,9648
	7	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9998	0,9993	0,9983	0,9961
9	0	0,6302	0,3874	0,2316	0,1342	0,0751	0,0404	0,0207	0,0101	0,0046	0,0020
	1	0,9288	0,7748	0,5995	0,4362	0,3003	0,1960	0,1211	0,0705	0,0385	0,0195
	2	0,9916	0,9470	0,8591	0,7382	0,6007	0,4628	0,3373	0,2318	0,1495	0,0898
	3	0,9994	0,9917	0,9661	0,9144	0,8343	0,7297	0,6089	0,4826	0,3614	0,2539
	4	1,0000	0,9991	0,9944	0,9804	0,9511	0,9012	0,8283	0,7334	0,6214	0,5000
	5	1,0000	0,9999	0,9994	0,9969	0,9900	0,9747	0,9464	0,9006	0,8342	0,7461
	6	1,0000	1,0000	1,0000	0,9997	0,9987	0,9957	0,9888	0,9750	0,9502	0,9102
	7	1,0000	1,0000	1,0000	1,0000	0,9999	0,9996	0,9986	0,9962	0,9909	0,9805
	8	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9997	0,9992	0,9980
10	0	0,5987	0,3487	0,1969	0,1074	0,0563	0,0282	0,0135	0,0060	0,0025	0,0010
	1	0,9139	0,7361	0,5443	0,3758	0,2440	0,1493	0,0860	0,0464	0,0232	0,0107
	2	0,9885	0,9298	0,8202	0,6778	0,5256	0,3828	0,2616	0,1673	0,0996	0,0547
	3	0,9990	0,9872	0,9500	0,8791	0,7759	0,6496	0,5138	0,3823	0,2660	0,1719
	4	0,9999	0,9984	0,9901	0,9672	0,9219	0,8497	0,7515	0,6331	0,5044	0,3770
	5	1,0000	0,9999	0,9986	0,9936	0,9803	0,9527	0,9051	0,8338	0,7384	0,6230
	6	1,0000	1,0000	0,9999	0,9991	0,9965	0,9894	0,9740	0,9452	0,8980	0,8281
	7	1,0000	1,0000	1,0000	0,9999	0,9996	0,9984	0,9952	0,9877	0,9726	0,9453
	8	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9995	0,9983	0,9955	0,9893
	9	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9997	0,9990
11	0	0,5688	0,3138	0,1673	0,0859	0,0422	0,0198	0,0088	0,0036	0,0014	0,0005
	1	0,8981	0,6974	0,4922	0,3221	0,1971	0,1130	0,0606	0,0302	0,0139	0,0059
	2	0,9848	0,9104	0,7788	0,6174	0,4552	0,3127	0,2001	0,1189	0,0652	0,0327
	3	0,9984	0,9815	0,9306	0,8389	0,7133	0,5696	0,4256	0,2963	0,1911	0,1133
	4	0,9999	0,9972	0,9841	0,9496	0,8854	0,7897	0,6683	0,5328	0,3971	0,2744
	5	1,0000	0,9997	0,9973	0,9883	0,9657	0,9218	0,8513	0,7535	0,6331	0,5000
	6	1,0000	1,0000	0,9997	0,9980	0,9924	0,9784	0,9499	0,9006	0,8262	0,7256
	7	1,0000	1,0000	1,0000	0,9998	0,9988	0,9957	0,9878	0,9707	0,9390	0,8867
	8	1,0000	1,0000	1,0000	1,0000	0,9999	0,9994	0,9980	0,9941	0,9852	0,9673
	9	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9998	0,9993	0,9978	0,9941
	10	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9998	0,9995
12	0	0,5404	0,2824	0,1422	0,0687	0,0317	0,0138	0,0057	0,0022	0,0008	0,0002
	1	0,8816	0,6590	0,4435	0,2749	0,1584	0,0850	0,0424	0,0196	0,0083	0,0032

Tabla D.5 Funciones de distribución binomiales (*continuación*)

		p									
n	***i***	**0,05**	**0,10**	**0,15**	**0,20**	**0,25**	**0,30**	**0,35**	**0,40**	**0,45**	**0,50**
	2	0,9804	0,8891	0,7358	0,5583	0,3907	0,2528	0,1513	0,0834	0,0421	0,0193
	3	0,9978	0,9744	0,9078	0,7946	0,6488	0,4925	0,3467	0,2253	0,1345	0,0730
	4	0,9998	0,9957	0,9761	0,9274	0,8424	0,7237	0,5833	0,4382	0,3044	0,1938
	5	1,0000	0,9995	0,9954	0,9806	0,9456	0,8822	0,7873	0,6652	0,5269	0,3872
	6	1,0000	0,9999	0,9993	0,9961	0,9857	0,9614	0,9154	0,8418	0,7393	0,6128
	7	1,0000	1,0000	0,9999	0,9994	0,9972	0,9905	0,9745	0,9427	0,8883	0,8062
	8	1,0000	1,0000	1,0000	0,9999	0,9996	0,9983	0,9944	0,9847	0,9644	0,9270
	9	1,0000	1,0000	1,0000	1,0000	1,0000	0,9998	0,9992	0,9972	0,9921	0,9807
	10	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9997	0,9989	0,9968
	11	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9998
13	0	0,5133	0,2542	0,1209	0,0550	0,0238	0,0097	0,0037	0,0013	0,0004	0,0001
	1	0,8646	0,6213	0,3983	0,2336	0,1267	0,0637	0,0296	0,0126	0,0049	0,0017
	2	0,9755	0,8661	0,6920	0,5017	0,3326	0,2025	0,1132	0,0579	0,0269	0,0112
	3	0,9969	0,9658	0,8820	0,7437	0,5843	0,4206	0,2783	0,1686	0,0929	0,0461
	4	0,9997	0,9935	0,9658	0,9009	0,7940	0,6543	0,5005	0,3530	0,2279	0,1334
	5	1,0000	0,9991	0,9925	0,9700	0,9198	0,8346	0,7159	0,5744	0,4268	0,2905
	6	1,0000	0,9999	0,9987	0,9930	0,9757	0,9376	0,8705	0,7712	0,6437	0,5000
	7	1,0000	1,0000	0,9998	0,9988	0,9944	0,9818	0,9538	0,9023	0,8212	0,7095
	8	1,0000	1,0000	1,0000	0,9998	0,9990	0,9960	0,9874	0,9679	0,9302	0,8666
	9	1,0000	1,0000	1,0000	1,0000	0,9999	0,9993	0,9975	0,9922	0,9797	0,9539
	10	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9997	0,9987	0,9959	0,9888
	11	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9995	0,9983
	12	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999
14	0	0,4877	0,2288	0,1028	0,0440	0,0178	0,0068	0,0024	0,0008	0,0002	0,0001
	1	0,8470	0,5846	0,3567	0,1979	0,1010	0,0475	0,0205	0,0081	0,0029	0,0009
	2	0,9699	0,8416	0,6479	0,4481	0,2811	0,1608	0,0839	0,0398	0,0170	0,0065
	3	0,9958	0,9559	0,8535	0,6982	0,5213	0,3552	0,2205	0,1243	0,0632	0,0287
	4	0,9996	0,9908	0,9533	0,8702	0,7415	0,5842	0,4227	0,2793	0,1672	0,0898
	5	1,0000	0,9985	0,9885	0,9561	0,8883	0,7805	0,6405	0,4859	0,3373	0,2120
	6	1,0000	0,9998	0,9978	0,9884	0,9617	0,9067	0,8164	0,6925	0,5461	0,3953
	7	1,0000	1,0000	0,9997	0,9976	0,9897	0,9685	0,9247	0,8499	0,7414	0,6074
	8	1,0000	1,0000	1,0000	0,9996	0,9978	0,9917	0,9757	0,9417	0,8811	0,7880
	9	1,0000	1,0000	1,0000	1,0000	0,9997	0,9983	0,9940	0,9825	0,9574	0,9102
	10	1,0000	1,0000	1,0000	1,0000	1,0000	0,9998	0,9989	0,9961	0,9886	0,9713
	11	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9994	0,9978	0,9935

Tabla D.5 Funciones de distribución binomiales (*continuación*)

		p									
n	***i***	**0,05**	**0,10**	**0,15**	**0,20**	**0,25**	**0,30**	**0,35**	**0,40**	**0,45**	**0,50**
	12	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9997	0,9991
	13	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999
15	0	0,4633	0,2059	0,0874	0,0352	0,0134	0,0047	0,0016	0,0005	0,0001	0,0000
	1	0,8290	0,5490	0,3186	0,1671	0,0802	0,0353	0,0142	0,0052	0,0017	0,0005
	2	0,9638	0,8159	0,6042	0,3980	0,2361	0,1268	0,0617	0,0271	0,0107	0,0037
	3	0,9945	0,9444	0,8227	0,6482	0,4613	0,2969	0,1727	0,0905	0,0424	0,0176
	4	0,9994	0,9873	0,9383	0,8358	0,6865	0,5155	0,3519	0,2173	0,1204	0,0592
	5	0,9999	0,9978	0,9832	0,9389	0,8516	0,7216	0,5643	0,4032	0,2608	0,1509
	6	1,0000	0,9997	0,9964	0,9819	0,9434	0,8689	0,7548	0,6098	0,4522	0,3036
	7	1,0000	1,0000	0,9996	0,9958	0,9827	0,9500	0,8868	0,7869	0,6535	0,5000
	8	1,0000	1,0000	0,9999	0,9992	0,9958	0,9848	0,9578	0,9050	0,8182	0,6964
	9	1,0000	1,0000	1,0000	0,9999	0,9992	0,9963	0,9876	0,9662	0,9231	0,8491
	10	1,0000	1,0000	1,0000	1,0000	0,9999	0,9993	0,9972	0,9907	0,9745	0,9408
	11	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9995	0,9981	0,9937	0,9824
	12	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9997	0,9989	0,9963
	13	1,0000	1,0000	1,9000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9995
	14	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
16	0	0,4401	0,1853	0,0743	0,0281	0,0100	0,0033	0,0010	0,0003	0,0001	0,0000
	1	0,8108	0,5147	0,2839	0,1407	0,0635	0,0261	0,0098	0,0033	0,0010	0,0003
	2	0,9571	0,7892	0,5614	0,3518	0,1971	0,0994	0,0451	0,0183	0,0066	0,0021
	3	0,9930	0,9316	0,7899	0,5981	0,4050	0,2459	0,1339	0,0651	0,0281	0,0106
	4	0,9991	0,9830	0,9209	0,7982	0,6302	0,4499	0,2892	0,1666	0,0853	0,0384
	5	0,9999	0,9967	0,9765	0,9183	0,8103	0,6598	0,4900	0,3288	0,1976	0,1051
	6	1,0000	0,9995	0,9944	0,9733	0,9204	0,8247	0,6881	0,5272	0,3660	0,2272
	7	1,0000	0,9999	0,9989	0,9930	0,9729	0,9256	0,8406	0,7161	0,5629	0,4018
	8	1,0000	1,0000	0,9998	0,9985	0,9925	0,9743	0,9329	0,8577	0,7441	0,5982
	9	1,0000	1,0000	1,0000	0,9998	0,9984	0,9929	0,9771	0,9417	0,8759	0,7728
	10	1,0000	1,0000	1,0000	1,0000	0,9997	0,9984	0,9938	0,9809	0,9514	0,8949
	11	1,0000	1,0000	1,0000	1,0000	1,0000	0,9997	0,9987	0,9951	0,9851	0,9616
	12	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9998	0,9991	0,9965	0,9894
	13	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9994	0,9979
	14	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9997
	15	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
17	0	0,4181	0,1668	0,0631	0,0225	0,0075	0,0023	0,0007	0,0002	0,0000	0,0000
	1	0,7922	0,4818	0,2525	0,1182	0,0501	0,0193	0,0067	0,0021	0,0006	0,0001

Tabla D.5 Funciones de distribución binomiales (*continuación*)

		p									
n	***i***	**0,05**	**0,10**	**0,15**	**0,20**	**0,25**	**0,30**	**0,35**	**0,40**	**0,45**	**0,50**
	2	0,9497	0,7618	0,5198	0,3096	0,1637	0,0774	0,0327	0,0123	0,0041	0,0012
	3	0,9912	0,9174	0,7556	0,5489	0,3530	0,2019	0,1028	0,0464	0,0184	0,0063
	4	0,9988	0,9779	0,9013	0,7582	0,5739	0,3887	0,2348	0,1260	0,0596	0,0245
	5	0,9999	0,9953	0,9681	0,8943	0,7653	0,5968	0,4197	0,2639	0,1471	0,0717
	6	1,0000	0,9992	0,9917	0,9623	0,8929	0,7752	0,6188	0,4478	0,2902	0,1662
	7	1,0000	0,9999	0,9983	0,9891	0,9598	0,8954	0,7872	0,6405	0,4743	0,3145
	8	1,0000	1,0000	0,9997	0,9974	0,9876	0,9597	0,9006	0,8011	0,6626	0,5000
	9	1,0000	1,0000	1,0000	0,9995	0,9969	0,9873	0,9617	0,9081	0,8166	0,6855
	10	1,0000	1,0000	1,0000	0,9999	0,9994	0,9968	0,9880	0,9652	0,9174	0,8338
	11	1,0000	1,0000	1,0000	1,0000	0,9999	0,9993	0,9970	0,9894	0,9699	0,9283
	12	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9994	0,9975	0,9914	0,9755
	13	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9995	0,9981	0,9936
	14	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9997	0,9988
	15	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999
	16	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
18	0	0,3972	0,1501	0,0536	0,0180	0,0056	0,0016	0,0004	0,0001	0,0000	0,0000
	1	0,7735	0,4503	0,2241	0,0991	0,0395	0,0142	0,0046	0,0013	0,0003	0,0001
	2	0,9419	0,7338	0,4797	0,2713	0,1353	0,0600	0,0236	0,0082	0,0025	0,0007
	3	0,9891	0,9018	0,7202	0,5010	0,3057	0,1646	0,0783	0,0328	0,0120	0,0038
	4	0,9985	0,9718	0,8794	0,7164	0,5187	0,3327	0,1886	0,0942	0,0411	0,0154
	5	0,9998	0,9936	0,9581	0,8671	0,7175	0,5344	0,3550	0,2088	0,1077	0,0481
	6	1,0000	0,9988	0,9882	0,9487	0,8610	0,7217	0,5491	0,3743	0,2258	0,1189
	7	1,0000	0,9998	0,9973	0,9837	0,9431	0,8593	0,7283	0,5634	0,3915	0,2403
	8	1,0000	1,0000	0,9995	0,9957	0,9807	0,9404	0,8609	0,7368	0,5778	0,4073
	9	1,0000	1,0000	0,9999	0,9991	0,9946	0,9790	0,9403	0,8653	0,7473	0,5927
	10	1,0000	1,0000	1,0000	0,9998	0,9988	0,9939	0,9788	0,9424	0,8720	0,7597
	11	1,0000	1,0000	1,0000	1,0000	0,9998	0,9986	0,9938	0,9797	0,9463	0,8811
	12	1,0000	1,0000	1,0000	1,0000	1,0000	0,9997	0,9986	0,9942	0,9817	0,9519
	13	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9997	0,9987	0,9951	0,9846
	14	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9998	0,9990	0,9962
	15	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9993
	16	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999
19	0	0,3774	0,1351	0,0456	0,0144	0,0042	0,0011	0,0003	0,0001	0,0000	0,0000
	1	0,7547	0,4203	0,1985	0,0829	0,0310	0,0104	0,0031	0,0008	0,0002	0,0000
	2	0,9335	0,7054	0,4413	0,2369	0,1113	0,0462	0,0170	0,0055	0,0015	0,0004

Tabla D.5 Funciones de distribución binomiales (*continuación*)

		p									
n	***i***	**0,05**	**0,10**	**0,15**	**0,20**	**0,25**	**0,30**	**0,35**	**0,40**	**0,45**	**0,50**
	3	0,9868	0,8850	0,6841	0,4551	0,2630	0,1332	0,0591	0,0230	0,0077	0,0022
	4	0,9980	0,9648	0,8556	0,6733	0,4654	0,2822	0,1500	0,0696	0,0280	0,0096
	5	0,9998	0,9914	0,9463	0,8369	0,6678	0,4739	0,2968	0,1629	0,0777	0,0318
	6	1,0000	0,9983	0,9837	0,9324	0,8251	0,6655	0,4812	0,3081	0,1727	0,0835
	7	1,0000	0,9997	0,9959	0,9767	0,9225	0,8180	0,6656	0,4878	0,3169	0,1796
	8	1,0000	1,0000	0,9992	0,9933	0,9713	0,9161	0,8145	0,6675	0,4940	0,3238
	9	1,0000	1,0000	0,9999	0,9984	0,9911	0,9674	0,9125	0,8139	0,6710	0,5000
	10	1,0000	1,0000	1,0000	0,9997	0,9977	0,9895	0,9653	0,9115	0,8159	0,6762
	11	1,0000	1,0000	1,0000	1,0000	0,9995	0,9972	0,9886	0,9648	0,9129	0,8204
	12	1,0000	1,0000	1,0000	1,0000	0,9999	0,9994	0,9969	0,9884	0,9658	0,9165
	13	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9993	0,9969	0,9891	0,9682
	14	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9994	0,9972	0,9904
	15	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9995	0,9978
	16	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9996
	17	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
20	0	0,3585	0,1216	0,0388	0,0115	0,0032	0,0008	0,0002	0,0000	0,0000	0,0000
	1	0,7358	0,3917	0,1756	0,0692	0,0243	0,0076	0,0021	0,0005	0,0001	0,0000
	2	0,9245	0,6769	0,4049	0,2061	0,0913	0,0355	0,0121	0,0036	0,0009	0,0002
	3	0,9841	0,8670	0,6477	0,4114	0,2252	0,1071	0,0444	0,0160	0,0049	0,0013
	4	0,9974	0,9568	0,8298	0,6296	0,4148	0,2375	0,1182	0,0510	0,0189	0,0059
	5	0,9997	0,9887	0,9327	0,8042	0,6172	0,4164	0,2454	0,1256	0,0553	0,0207
	6	1,0000	0,9976	0,9781	0,9133	0,7858	0,6080	0,4166	0,2500	0,1299	0,0577
	7	1,0000	0,9996	0,9941	0,9679	0,8982	0,7723	0,6010	0,4159	0,2520	0,1316
	8	1,0000	0,9999	0,9987	0,9900	0,9591	0,8867	0,7624	0,5956	0,4143	0,2517
	9	1,0000	1,0000	0,9998	0,9974	0,9861	0,9520	0,8782	0,7553	0,5914	0,4119
	10	1,0000	1,0000	1,0000	0,9994	0,9961	0,9829	0,9468	0,8725	0,7507	0,5881
	11	1,0000	1,0000	1,0000	0,9999	0,9991	0,9949	0,9804	0,9435	0,8692	0,7483
	12	1,0000	1,0000	1,0000	1,0000	0,9998	0,9987	0,9940	0,9790	0,9420	0,8684
	13	1,0000	1,0000	1,0000	1,0000	1,0000	0,9997	0,9985	0,9935	0,9786	0,9423
	14	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9997	0,9984	0,9936	0,9793
	15	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9997	0,9985	0,9941
	16	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9997	0,9987
	17	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9998
	18	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

Programas

Programa	Lo que computa
3-1	Media muestral, varianza muestral, desviación típica muestral
3-2	Coeficiente de correlación muestral
5-1	Probabilidades binomiales
6-1	Distribución de probabilidad normal estándar
6-2	Percentiles de la distribución normal estándar
8-	Percentiles de las distribuciones *t*
8-2	Probabilidades de las distribuciones *t*
8-3	Estimadores por intervalo de confianza y cotas de confianza para la media
9-1	*p* valor para los contrastes de la *t*
10-1	*p* valor para los contrastes de la *t* con dos muestras
11-1	*p* valor para el ANOVA unifactorial
11-2	*p* valor para el ANOVA con dos factores
12-1	Estadísticos en el modelo de regresión lineal simple
12-2	Estimadores por mínimos cuadrados en la regresión lineal múltiple
13-1	*p* valor en los contrastes de bondad de ajuste de la chi-cuadrado
13-2	*p* valor en los contrastes de independencia en tablas de contingencia
14-1	*p* valor en los contrastes de rangos signados
14-2	*p* valor en los contrastes de suma de rangos
14-3	*p* valor en los contrastes de rachas
A-1	Subconjuntos aleatorios

Respuestas a los problemas con número impar

Problemas del capítulo 1

1. (a) 1946

   (b) Hay más años en los que el promedio de cursos completados por el grupo de mayor edad fue superior al del grupo más joven.

3. (a) Desde 1985 a 1990 disminuyeron las ventas.

   (b) Entre 1985 y 1987 se vendieron 20 693 000 coches, frente a los 18 120 000 vendidos entre 1988 y 1990.

   (c) No

5. Los investigadores que lo supieran podrían estar influenciados por sus propios sesgos respecto a la utilidad del medicamento.

7. (a) En 1936, los propietarios de coches y los titulares de teléfonos probablemente no eran representativos de la población total de votantes.

   (b) Sí. En la actualidad, la propiedad de coches y la titularidad de teléfonos están muy extendidas y, por consiguiente, sí que sería una muestra más representativa de la población total de votantes.

9. La edad media de muerte de los ciudadanos de Estados Unidos, cuyo obituario aparece en la lista de *The New York Times*, es de 82,4 años.

11. (a) No. Es posible que los licenciados que reenvían el cuestionario no sean representativos de la población total de licenciados.

    (b) Si el número de cuestionarios devueltos se aproximara a los 200 –el número de cuestionarios enviados–, la aproximación sería mejor.

13. Graunt asumió implícitamente que las parroquias muestreadas eran representativas de la población total de Londres.

15. Se pueden utilizar los datos sobre las edades de muerte para calcular aproximadamente el número medio de años en los que se continuarán realizando los pagos de las anualidades. Este número medio se podría utilizar después para determinar el cargo.

17. (a) 64%

(b) 10%

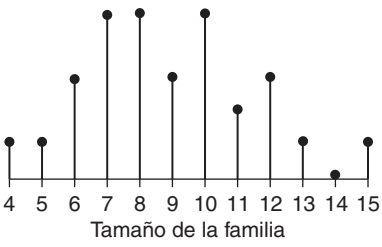
(c) 48%

Sección 2.2

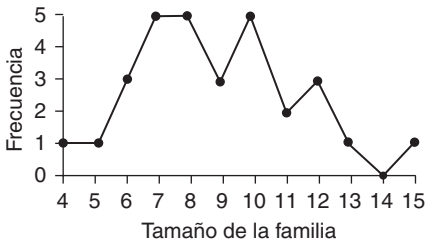
1. (a)

Tamaño de la familia	Frecuencia
4	1
5	1
6	3
7	5
8	5
9	3
10	5
11	2
12	3
13	1
14	0
15	1

(b)



(c)



3. (a) 12

   (b) 1

   (c) 11

   (d) 3

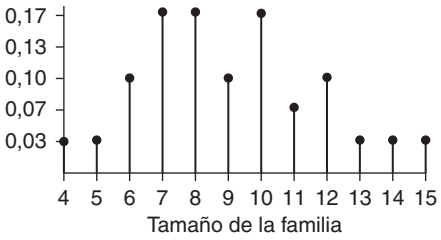
   (e) 3

5.

Valor	Frecuencia
10	8
20	3
30	7
40	7
50	3
60	8

7.

Tamaño de la familia	Frecuencia	Frecuencia relativa
4	1	0,03
5	1	0,03
6	3	0,10
7	5	0,17
8	5	0,17
9	3	0,10
10	5	0,17
11	2	0,07
12	3	0,10
13	1	0,03
14	0	0,00
15	1	0,03



9. (a) 0,13

   (b) 0,25

   (c) No

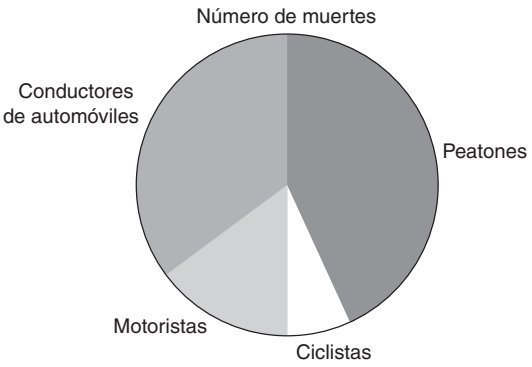
11. (a) 0,649

(b) 0,162

(c) 0,540

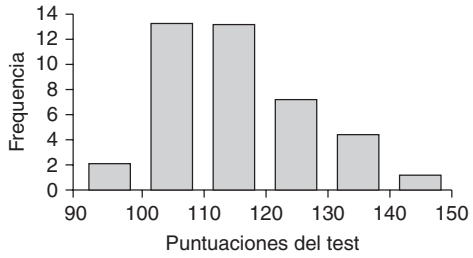
13.

Número medio de días de lluvia en nov. o dic.	Frecuencia
7	1
9	1
10	1
11	1
16	1
17	3
18	1
20	1
23	1
40	1



Sección 2.3

1. (a)

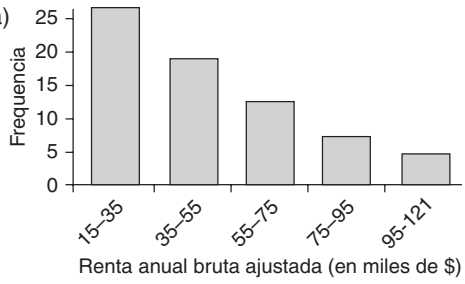


(b) Los intervalos de clase son 100-110 y 110-120.

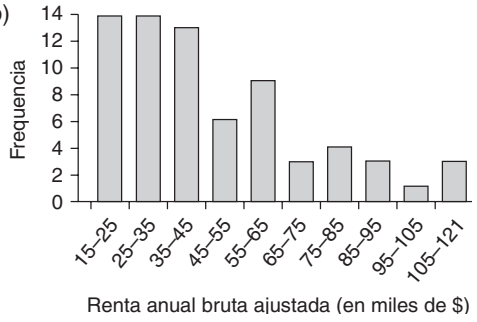
(c) No

(d) No

3. (a)

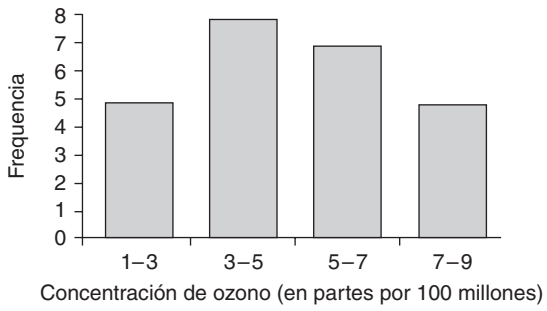


(b)

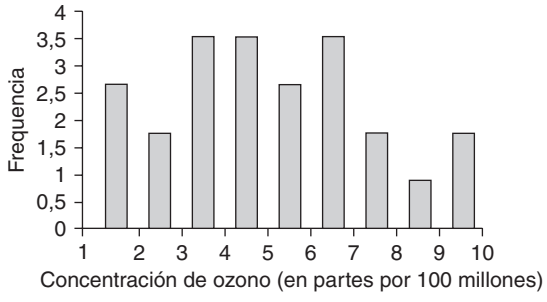


(c) El histograma de la parte (a) parece más informativo, puesto que refleja un patrón más claro.

5. (a)

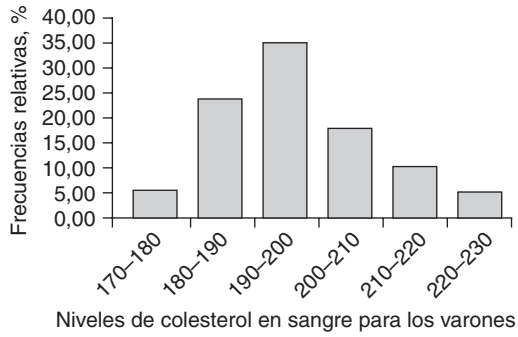


(b)

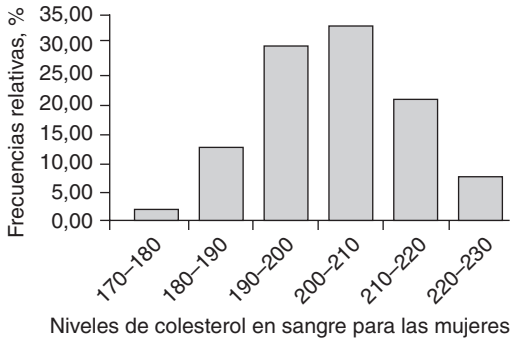


(c) El histograma del apartado (b) parece más informativo.

7. (a)

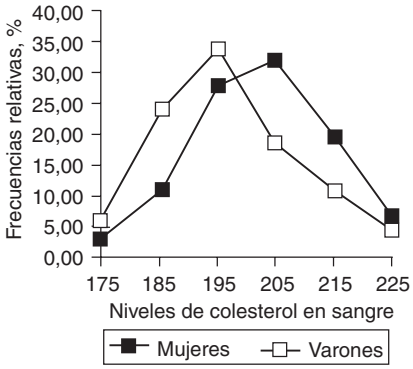


(b)



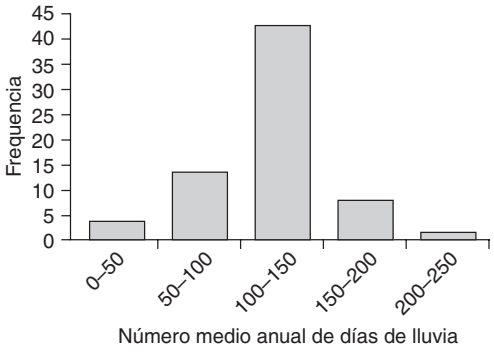
Colesterol en mujeres	Frecuencias	Frecuencias relativas
170–180	1	1/46 = 0,02
180–190	5	5/46 = 0,11
190–200	13	13/46 = 0,28
200–210	15	15/46 = 0,33
210–220	9	9/46 = 0,20
220–230	3	3/46 = 0,07

Colesterol en varones	Frecuencias	Frecuencias relativas
170–180	3	3/54 = 0,06
180–190	13	13/54 = 0,24
190–200	19	19/54 = 0,35
200–210	10	10/54 = 0,19
210–220	6	6/54 = 0,11
220–230	3	3/54 = 0,06



Entre los estudiantes, parece que las mujeres tengan niveles de colesterol más altos.

13.



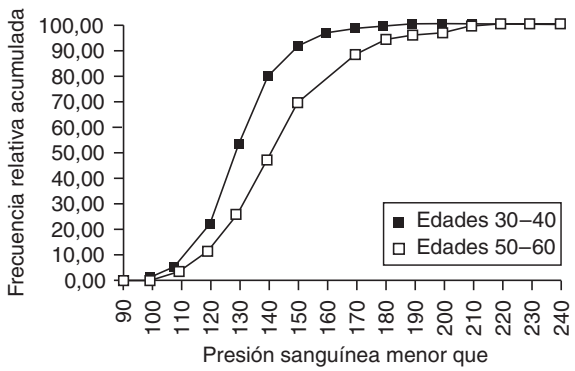
15. (a) Es la suma de las frecuencias relativas de todas las clases.

(b)

Presión sanguínea menor que	Porcentaje de trabajadores	
	Con edad de 30-40	Con edad de 50-60
90	0,12	0,14
100	0,79	0,41
110	5,43	3,56
120	23,54	11,35
130	53,78	28,04
140	80,35	48,43
150	92,64	71,27
160	97,36	81,26
170	99,13	89,74
180	99,84	94,53
190	99,96	97,26
200	100,00	98,50
210	100,00	98,91
220	100,00	99,59
230	100,00	99,86
240	100,00	100,00

(c) Los trabajadores con una edad comprendida entre 30 y 40 años tienden a tener valores más bajos de presión sanguínea.

(d)



Sección 2.4

1. (a)

11	1, 4, 5, 6, 8, 8, 9, 9, 9
12	2, 2, 2, 2, 4, 5, 5, 6, 7, 7, 7, 8, 9
13	0, 2, 2, 3, 4, 5, 5, 7, 9
14	1, 1, 4, 6, 7

(b

11	1, 4
11	5, 6, 8, 8, 9, 9, 9
12	2, 2, 2, 2, 4
12	5, 5, 6, 7, 7, 7, 8, 9
13	0, 2, 2, 3, 4
13	5, 5, 7, 9
14	1, 1, 4
14	6, 7

3.

1	4
1	5, 6, 6, 7, 7, 7, 7, 8, 8, 8, 9, 9, 9, 9
2	0, 0, 0, 0, 1, 2, 2, 2, 3, 4
2	5, 7, 7, 9
3	0, 1, 1, 2, 3
3	
4	0, 4, 4
4	5
5	1, 3
5	5
6	1
6	
7	
7	9

El intervalo 15-20 contiene 14 valores de datos.
El intervalo 16-21 contiene 17 valores de datos.

5. (a)

3	2
4	
5	2, 7, 8, 9
6	5, 8, 8
7	1, 4, 5, 5, 7, 8, 9
8	0, 1, 3, 3, 3, 4, 8, 8
9	0, 3, 4, 7
10	0, 4, 8

(b) Sí. El valor 32 parece sospechoso puesto que es mucho menor que los restantes.

7. (a)

1	4, 6, 6, 6
2	0, 0, 1, 3, 4, 4, 6, 7, 7, 7
3	1, 2, 3, 5, 5, 8, 8, 9
4	2, 6
5	5

(b)

0	3, 6, 7, 7, 7, 7, 9
1	0, 0, 0, 0, 0, 0, 3, 4, 4, 6, 6, 7, 7, 9, 9
2	0, 1
3	1

(c)

0	1, 3, 4, 4, 4, 5, 7, 9
1	0, 0, 2, 6, 7, 7, 7, 8, 9, 9
2	1, 2, 5, 9
3	2, 6
4	5

9. (a) 6

(b) 43,75%

(c) 12,5%

11. (a) Escuela B

(b) Escuela A

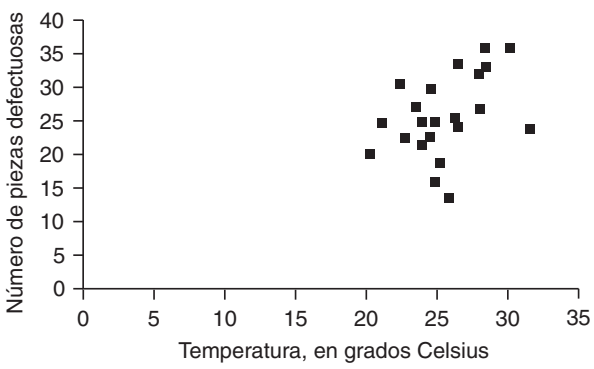
(c) Escuela A

(d)

5	0, 3
5	5, 7
6	2
6	5, 5, 8, 8, 9, 9
7	0, 2, 3, 4
7	6, 7, 7, 8, 8, 9, 9
8	0, 2, 3, 3
8	5, 5, 6, 6, 6, 7, 7, 8, 8, 9
9	0, 0, 1, 3
9	5, 5, 5, 6, 6, 8, 8
10	0

Sección 2.5

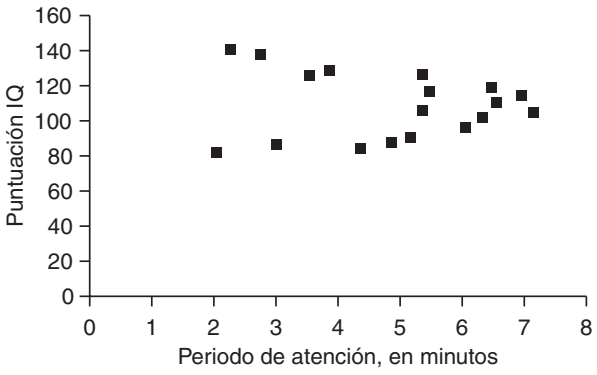
1. (a)



(b) El número de piezas defectuosas tiende a crecer a medida que aumenta la temperatura.

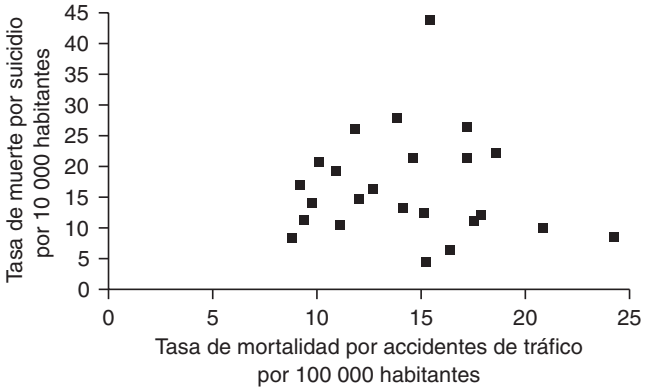
(c) Estarán entre 23 y 24.

5. (a)



(b) El periodo de atención y la puntuación IQ no están relacionados.

9. Las dos causas parecen no estar relacionadas



Capítulo 2 Problemas de repaso

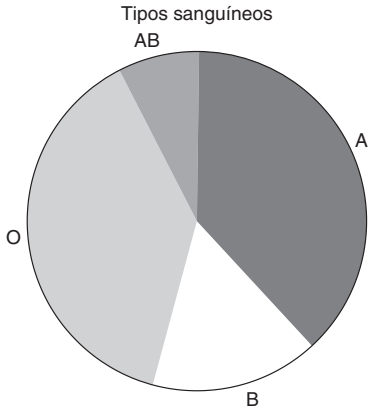
1. (a)

Tipo sanguíneo	Frecuencia
A	19
B	8
O	19
AB	4

(b)

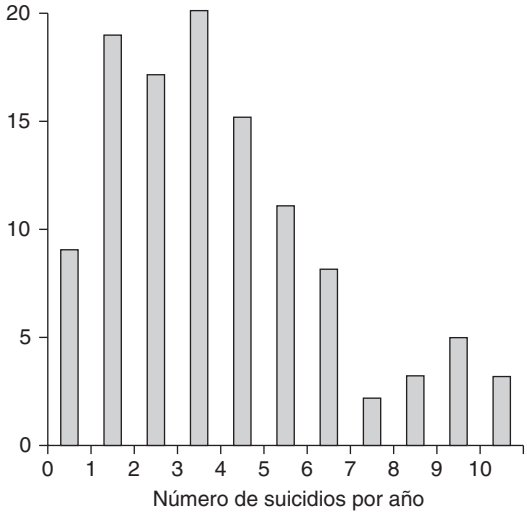
Tipo sanguíneo	Frecuencia relativa
A	0,38
B	0,16
O	0,38
AB	0,08

(c)



3. (a) 389

(b)



5. (a)

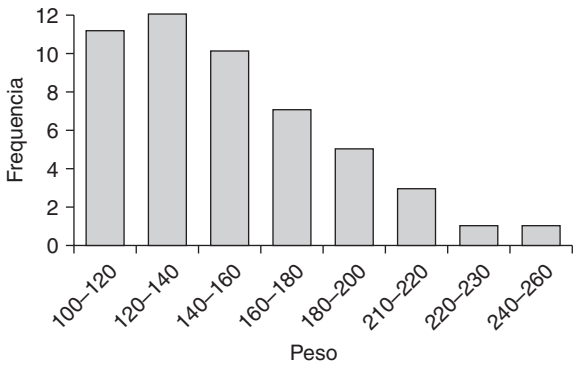
Valores	Frecuencias
1	2
2	1
3	4
4	1
5	2

(b)

Valores	Frecuencias
1	2
2	3
3	3
4	2

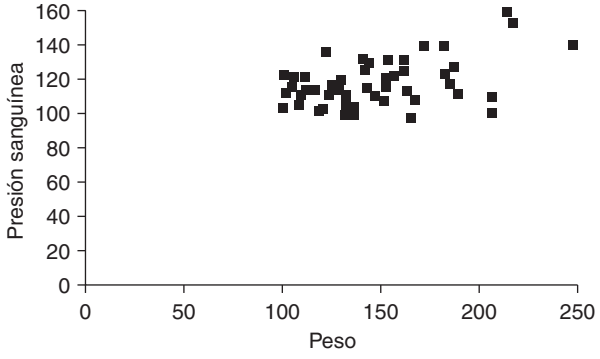
(c) 3, 2,5

9. (a)

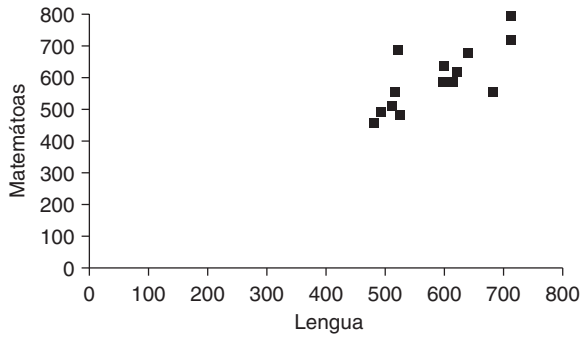


(b) Relativamente existen pocos pesos dentro del rango de pesos más altos.

11. El peso y la presión sanguínea no parecen estar relacionados.



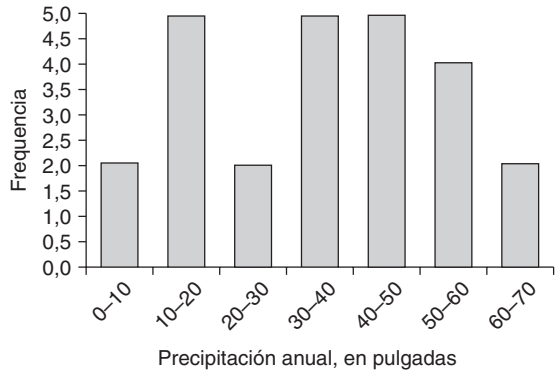
13. Las calificaciones altas en una asignatura tienden a aparecer junto a las calificaciones altas en la otra.



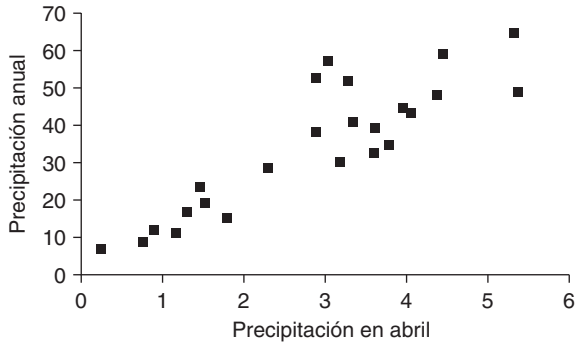
15. (a)

0	0,27, 0,78, 0,93
1	0,19, 0,31, 0,49, 0,53, 0,81
2	0,30, 0,92, 0,93
3	0,07, 0,21, 0,32, 0,39, 0,66, 0,68, 0,81
4	0,02, 0,11, 0,43, 0,50
5	0,35, 0,41

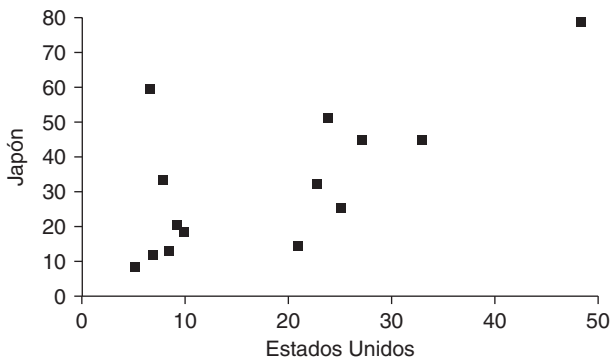
(b)



(c)



19.



Sección 3.2

1. 1196/15 = 79,73
3. 429,03/13 = 33,00 pulgadas; 1331/13 = 102,38 días
5. 1807/12 = 150,58
7. 638/7 = 91,14 casos
9. 6; 18; 11
11. 78/11
13. 15
15. $\frac{1}{2}(10) + \frac{1}{6}(20) + \frac{1}{3}(30) = 18{,}33$
17. 37 120 $.
19. (a) $-5, -4, -2, 1, 4, 6$
    (b) $-15, -12, -6, 3, 12, 18$
    (c) Igual que en (a).

Sección 3.3

1. (a) 6580 yardas
   (b) 6545 yardas
3. 23
5. (a) 22,0
   (b) 8,1
   (c) 23,68
   (d) 9,68
7. 31,5 pulgadas.

9. (a) 99,4
   (b) 14,9
   (c) 204,55
11. (a) 20,74
   (b) 20,5
   (c) 19,74
   (d) 19,5
   (e) Media = 20,21; mediana = 20,05
13. 0, 0
15. (a) 32,52
   (b) 24,25
17. (a) 26,8
   (b) 25,0

Sección 3.3.1

1. (a) Si los datos están ordenados de menor a mayor, el percentil de orden 80% coincide con la media de los valores que ocupan las posiciones 60 y 61.
   (b) Si los datos están ordenados en forma creciente, el percentil de orden 60% viene dado por la media de los valores que ocupan las posiciones 45 y 46.
   (c) Si los datos están ordenados de menor a mayor, el percentil de orden 30% es igual al valor que ocupa la posición 23.
3. (a) 95,5
   (b) 96
5. (a) 70
   (b) 58
   (c) 52
7. 230*c*
11. 25

Sección 3.4

1. 1*B*, 2*C*, 3*A*
3. (a) 126
   (b) 102, 110, 114
   (c) 196

5. 5, 6, 6, 6, 8, 10, 12, 14, 23 es un conjunto de datos que cumple las condiciones pedidas.

7. (a) 8 vueltas.

   (b) 2 millas.

Sección 3.5

1. $s^2 = 0{,}037;\ \bar{x} = 26{,}22$

3. (a) 6,18

   (b) 6,77

11. (a) $s^2 = 2{,}5,\ s = 1{,}58$

    (b) $s^2 = 2{,}5,\ s = 1{,}58$

    (c) $s^2 = 2{,}5,\ s = 1{,}58$

    (d) $s^2 = 10,\ s = 3{,}16$

    (e) $s^2 = 250,\ s = 15{,}81$

13. Para los primeros 50 estudiantes, $s^2 = 172{,}24$ y $\bar{x} = 115{,}80$.
    Para los 50 últimos estudiantes, $s^2 = 178{,}96$ y $\bar{x} = 120{,}98$.
    Los valores de los estadísticos son similares para los dos conjuntos de datos. Esto no es sorprendente.

15. 195 808,76

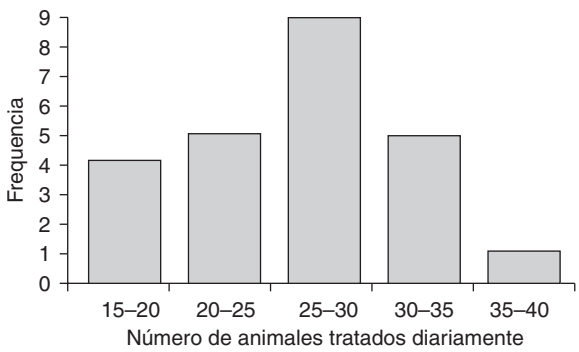
17. (a) 0,805

    (b) 2,77

    (c) 1,22

Sección 3.6

1. (a)



   (b) 25,75

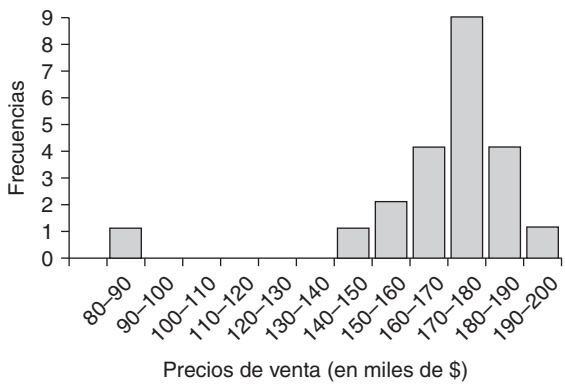
(c) 26,5

(d) No

5. (a) 168 045

(b) 172 500

(c)



(d) Sí, si se ignoran los 82 valores de datos. No, si se usan todos los datos.

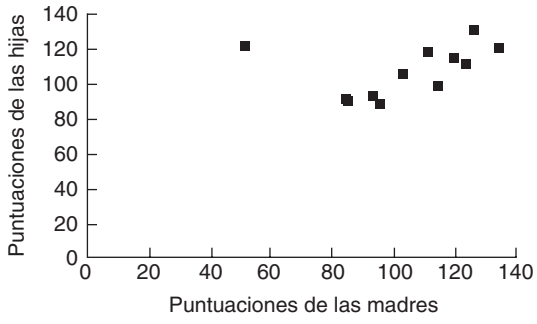
7. 95%, 94,85%

9. La media muestral

Sección 3.7

1. Si se denota por (x_i, y_i), $i = 1, 2, 3$, el conjunto central de pares de datos, el primer conjunto de datos es $(121x_i, 360 + y_i)$ y el tercero es $(x_i, \frac{1}{2}y_i)$, $i = 1, 2, 3$.

3. (a)



(b) Casi 1

(c) 0,86

(d) Entre ellas existe una relación lineal relativamente fuerte.

5. $-0{,}59$; la relación lineal es relativamente débil.

7. −0,441202; la relación lineal es relativamente débil. Sin embargo, existe una cierta indicación de que, cuando una de las variables es alta, la otra tiende a ser baja.

9. 0,95

11. Si se usan todos los datos, el coeficiente de correlación muestral es igual a −0,33; si se utilizan los primeros siete países, el coeficiente es −0,046.

13. Si se usan todos los datos, el coeficiente de correlación muestral es igual a 0,25; si se utilizan los primeros siete países, el coeficiente es −0,3035.

15. (d) Correlación no implica causalidad.

17. No, correlación no es causalidad.

Capítulo 3 Problemas de repaso

1. (a) −2, −1, 1, 2

   (b) −2, −1, 0, 1, 2

   (c) Parte (a): media = 0, mediana = 0; parte (b): media = 0, mediana = 0

3. (a) 29,3

   (b) No

   (c) El primer cuartil es 27,7; el segundo cuartil, 29,3; el tercer cuartil, 31,1

   (d) 31,7

9. No

11. No, asociación no es causalidad.

Sección 4.2

1. (a) $S = \{(R, R), (R, B), (R, Y), (B, R), (B, B), (B, Y), (Y, R), (Y, B), (Y, Y)\}$

   (b) $\{(Y, R), (Y, B), (Y, Y)\}$

   (c) $\{(R, R), (B, B), (Y, Y)\}$

3. (a) {(Michigan, Oregón), (Michigan, San José), (Reed, Oregón), (Reed, San José), (San José, Oregón), (San José, San José), (Yale, Oregón), (Yale, San José), (Oregón, Oregón), (Oregón, San José)}

   (b) {(San José, San José), (Oregón, Oregón)}

   (c) {(Michigan, Oregón), (Michigan, San José), (Reed, Oregón), (Reed, San José), (Yale, Oregón), (Yale, San José), (Oregón, San José)}

   (d) {(Reed, Oregón), (Oregón, Oregón), (San José, San José)}

5. S = {(Reed, Oregón), (Oregón, Oregón), (San José, San José)}
   A = {(Francia, avión), (Canadá, avión)}

7. (a) $\emptyset$

   (b) $\{1, 4, 6\}$

   (c) $\{1, 3, 4, 5\}$

   (d) $\{2\}$

9. (a) $\{(1, g), (1, f), (1, s), (1, c), (0, g), (0, f), (0, s), (0, c)\}$

   (b) $\{(0, s), (0, c)\}$

   (c) $\{(1, g), (1, f), (0, g), (0, f)\}$

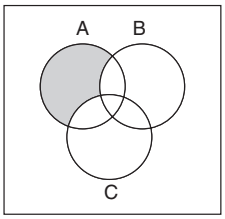
   (d) $\{(1, g), (1, f), (1, s), (1, c)\}$

11. (a) A^c es el suceso que consiste en obtener un número impar en el lanzamiento del dado.

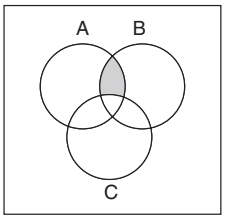
    (b) $(A^c)^c$ es el suceso que consiste en obtener un número par al lanzar el dado.

    (c) $(A^c)^c = A$.

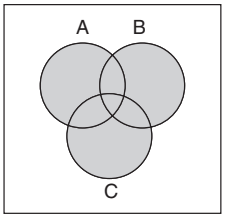
13. (a)



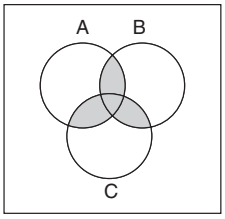
(b)



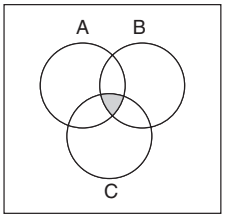
(c)



(d)



(e)



Sección 4.3

1. (a) $P(E) = 0{,}35$; $P(E) = 0{,}65$; $P(G) = 0{,}55$

   (b) $P(E \cup F) = 1$

   (c) $P(E \cup G) = 0{,}8$

(d) $P(F \cup G) = 0{,}75$

(e) $P(E \cup F \cup G) = 1$

(f) $P(E \cap F) = 0$

(g) $P(F \cap G) = 0{,}45$

(h) $P(E \cap G) = 0{,}1$

(i) $P(E \cap F \cap G) = 0$

3. 1/10 000

5. Si son disjuntos, es imposible. Si no son disjuntos, es posible.

7. (a) 1

(b) 0,8

(c) 0,5

(d) 0,1

9. (a) 0,95

(b) 0,80

(c) 0,20

11. 0,7

13. 0,31%

15. 0,6

17. (a) $A \cap B^c$

(b) $A \cap B$

(c) $B \cap A^c$

(d) $P(\text{I}) + P(\text{II}) + P(\text{III})$

(e) $P(\text{I}) + P(\text{II})$

(f) $P(\text{II}) + P(\text{III})$

(g) $P(\text{II})$

Sección 4.4

1. $88/216 \approx 0{,}41$

3. (a) $4/52 \approx 0{,}08$

(b) $48/52 \approx 0{,}92$

(c) $13/52 \approx 0{,}25$

(d) $1/52 \approx 0{,}02$

5. (a) 78%
   (b) 69,9%
   (c) 24,3%
   (d) 61%
   (e) 87%
7. (a) 0,56
   (b) 0,1
9. (a) 0,4
   (b) 0,1
11. 56
13. 1/19
15. (a) 0,1
   (b) 0,1
17. (a) 10/31
   (b) 9/31
   (c) 1/3
   (d) 11/31
   (e) 7/31

Sección 4.5

1. (a) $0{,}02/0{,}3 \approx 0{,}067$
   (b) $0{,}02/0{,}03 \approx 0{,}667$
3. (a) 0,245
   (b) 0,293
5. (a) 0,145
   (b) 0,176
   (c) 0,215
   (d) 0,152
7. (a) 0,46
   (b) 0,65

9. (a) 262/682

   (b) 262/682

   (c) 350/682

   (d) 602/682

   (e) 598/682

   (f) 519/682

11. $1/169 \approx 0{,}006$

13. 0,6960

15. (a) $19/34 \approx 0{,}56$

    (b) $1 - 19/34 \approx 0{,}44$

    (c) $1/17 \approx 0{,}06$

17. Puesto que $P(B|A) > P(B)$, $P(A \cap B) > P(B)P(A)$

    Por consiguiente, $P(A|B) = \dfrac{P(A \cap B)}{P(B)} > \dfrac{P(B)P(A)}{P(B)} = P(A)$

19. 0,24

21. 0,68

23. (a) $7/12 \approx 0{,}58$

    (b) 50

    (c) $13/119 \approx 0{,}11$

    (d) $35/204 \approx 0{,}17$

    (e) 0,338

25. (a) 0,79; 0,21

    (b) 0,81; 0,27

27. (a) 1/2

    (b) 3/8

    (c) 2/3

29. 1/16

31. No; si los amigos no se conocieran uno a otro.

33. $P(A) = 1/13$; $P(B) = 1/4$; $P(A \cap B) = 1/52$; así pues $P(A \cap B) = P(A)P(B)$.

35. 1/365

37. (a) 0,64
    (b) 0,96
    (c) 0,8704
39. Sí, $P(A)P(B) = P(A \cap B)$
41. (a) $32/4805 \approx 0{,}0067$
    (b) $729/1922 \approx 0{,}38$
    (c) 0,060
    (d) 0,045
    (e) 0,006
    (f) 0,111
43. (a) 1/4
    (b) 2/3

Sección 4.6

1. (a) 0,55
   (b) 5/9
3. (a) 0,672
   (b) 0,893
5. 0,398
7. (a) 0,534
   (b) 0,402

Capítulo 4 Problemas de repaso

1. (a) 3/4
   (b) 3/4
   (c) 6/11
   (d) 1/22
   (e) 9/22
3. (a) 0,68
   (b) 0,06
   (c) 0,12

5. (a) 11/24

   (b) 13/23

7. (a) 1/64

   (b) 1/64

   (c) 1/64

9. (a) S = {(pollo, arroz, melón), (pollo, arroz, helado), (pollo, arroz, gelatina), (pollo, patatas, melón), (pollo, patatas, melón), (pollo, patatas, helado), (pollo, patatas, gelatina), (filete, arroz, melón), (filete, arroz, helado), (filete, arroz, gelatina), (filete, patatas, melón), (filete, patatas, helado), (filete, patatas, gelatina)}

   (b) {(pollo, patatas, helado), (pollo, patatas, gelatina), (filete, patatas, helado), (filete, patatas, gelatina)}

   (c) 1/3

   (d) 1/12

11. (a) 1/3

    (b) 1/3

    (c) 1/3

    (d) 1/2

13. 14/33 $\approx$ 0,424

15. (a) 1/52

    (b) 1/52

    (c) Igualmente

    (d) 1/52

17. (a) 0,42

    (b) 0,18

    (c) 0,24

    (d) 0,58

    (e) 0,724

19. No

21. (a) 0,496

    (b) 54/252

    (c) 36/248

    (d) No

23. (a) 4

(b)(i) 4/86

(b)(ii) 1/2

(b)(iii) No

25. (a) 0,077

(b) 0,0494

(c) 0,0285

Sección 5.2

1. $P\{Y = 0\} = 1/4$

$P\{Y = 1\} = 3/4$

3. (a) 5/12

(b) 5/12

(c) 0

(d) 1/4

5.

i	$P\{Y = i\}$
1	11/36
2	1/4
3	7/36
4	5/36
5	1/12
6	1/36

7.

i	$P\{X = i\}$
2	0,58
3	0,42

9.

i	$P\{X = i\}$
0	1199/1428
1	55/357
2	3/476

11.

i	$P\{X = i\}$
0	0,075
1	0,325
2	0,6

13. No; $P(4)$ es negativo.

15.

i	$P\{X = i\}$
0	38/223
1	82/223
2	57/223
3	34/223
4	10/223
5	2/223

17. (a) 0,1

    (b) 0,5

19.

i	$P\{X = i\}$
0	0,30
1	0,35
2	0,20
3	0,15

21. (a) 0,0711

    (b) 0,0018

23.

i	$P\{X = i\}$
0	0,855
100 000	0,14
200 000	0,005

Sección 5.3

1. (a) 2

   (b) 5/3

   (c) 7/3

3. 8,40 $

5. 1,9

7. (a) 2,53

   (b) 4,47

9. 880 $

11. (a) 2/3
    (b) 4/3
    (c) 2
13. (a) Segunda dirección
    (b) Primera dirección
15. −5 $
17. (a) No
    (b) No
    (c) Sí
    (d) $4/95 \approx 0{,}042$
19. −0,40 $
21. 2,5
23. 150 $
25. 0
27. (a) 16 800 $
    (b) 18 000 $
    (c) 18 000 $
29. 3
31. (a) 7
    (b) 7
33. 12
35. 3,6

Sección 5.4

1. $\text{Var}(U) = 0, \text{Var}(V) = 1, \text{Var}(W) = 100$
3. 0
5. 0,49
7. 0,25
9. (b) 0,8
   (c) 0,6

11. (a) 0,5

    (b) 0,5

13. (a) 0

    (b) 3666 $

15. (a) 4,06

    (b) 1,08

17. 3 SD(X) = 6

19. (a) 2

    (b) 2

Sección 5.5

1. (a) 24

   (b) 120

   (c) 5040

3. 3 628 800

5. (a) 0,278692

   (b) 0,123863

   (c) 0,00786432

7. (a) 0,468559

   (b) 0,885735

9. (a) 3 o más

   (b) 0,00856

11. 0,144531

13. (a) 0,517747

    (b) 0,385802

    (c) 0,131944

15. (a) 0,421875

    (b) 0,421875

    (c) 0,140625

    (d) 0,015625

17. (a) 10/3
    (b) 20/3
    (c) 10
    (d) 50/3
19. (a) 0,430467
    (b) 0,382638
    (c) 7,2
    (d) 0,72
21. (a) 0,037481
    (b) 0,098345
    (c) 0,592571
    (d) 1,76
    (e) 0,992774
23. (a) 0,00604662
    (b) 0
25. (a) 50; 5
    (b) 40; 4,89898
    (c) 60; 4,89898
    (d) 25; 3,53553
    (e) 75; 6,12372
    (f) 50; 6,12372

Sección 5.6

1. Hipergeométrica, $n = 20$, $N = 200$, $p = 0{,}09$
3. Hipergeométrica, $n = 6$, $N = 54$, $p = 6/54$
5. Hipergeométrica, $n = 20$, $N = 100$, $p = 0{,}05$
7. Binomial, $n = 10$, $p = 1/13$

Capítulo 5 Problemas de repaso

1. (a) 0,4
    (b) 0,6

3. (a) 1, 2, 3, 4

(b)

i	$P(X = i)$
1	0,3
2	0,21
3	0,147
4	0,343

(c) 0,7599

(d) 2,53

(e) 1,53

5. (a) 0,723

(b) No, porque si ella gana, ganará 1 $, mientras que si pierde, perderá 3 $.

(c) −0,108

7. (a)

i	$P(X = i)$
0	0,7
4000	0,15
6000	0,15

(b) 1500

(c) 5 550 000

(d) 2 355 84

9. La oferta a la baja maximiza el beneficio esperado.

11. (a) 1/3

(b) 1/4

(c) 7/24

(d) 1/12

(e) 1/24

(f) 625 $

(g) 125 $

13. (a) 0

(b) −68,750

(c) −68,750

17. (a) 0,6
    (b) 0,648
    (c) 0,68256
    (d) 0,710208
    (e) 0,733432
    (f) 0,813908
19. (a) 0,064
    (b) 0,432
    (c) 0,820026

Sección 6.2

1. (a) 0,29
   (b) 0,56
   (c) 0,33
   (d) 0,27
3. (a) 2/3
   (b) 0,7
   (c) 0,6
   (d) 0,6
5. (a) 2/3
   (b) 1/6
   (c) 1/3
7. (a) 1/2
   (b) 0
   (c) 3/4
   (d) 3/8

Sección 6.3

1. (a) De 108,8 a 148
   (b) De 89,2 a 167,6
   (c) De 69,6 a 187,2

3. (b)
5. (d)
7. (c)
9. (a)
11. (b)
13. (d)
15. (b)
17. (a) Y
    (b) X
    (c) X e Y tienen la misma probabilidad de ser mayores que 100.
19. (a) No
    (b) No
    (c) No
    (d) Sí

Sección 6.4

1. (a) 0,9861
   (b) 0,1357
   (c) 0,4772
   (d) 0,7007
   (e) 0,975
   (f) 0,2358
   (g) 0,899
   (h) 0,2302
   (i) 0,8710
3. 3
7. (a) 1,65
   (b) 1,96
   (c) 2,58
   (d) 0
   (e) 0,41

(f) 2,58

(g) 1,15

(h) 0,13

(i) 0,67

Sección 6.6

1. Puesto que $x > a$, también $x - u > a - u$. Así pues, $\frac{x - \mu}{\sigma} > \frac{a - \mu}{\sigma}$ ya que σ es positivo.
3. 0,3085
5. (a) 0,6179
   (b) 0,8289
   (c) 0,4468
7. 0,008
9. (a) 0,1587
   (b) 0,2514
   (c) 0,4772
11. 0,8664
13. (a) 0,0993
   (b) 0,011
   (c) 0,8996
15. (a) 0,2660
   (b) 0,9890
   (c) 0,7230
   (d) 0,9991
   (e) 0,0384

Sección 6.7

1. (a) 1,48
   (b) 1,17
   (c) 0,52
   (d) 1,88
   (e) −0,39
   (f) 0

(g) −1,64

(h) 2,41

3. (a) 50

(b) 57,68

(c) 61,76

(d) 40,16

(e) 57,02

5. 464,22

7. 525,6

9. 746

11. (a) Cierta

(b) Cierta

13. 99,28

Capítulo 6 Problemas de repaso

1. (a) 0,9236

(b) 0,8515

(c) 0,0324

(d) 0,9676

(e) 0,1423

(f) 0,0007

(g) 75,524

(h) 73,592

(i) 68,3

3. 4,969

5. (a) 0,1587

(b) 0,1587

(c) 0,1886

(d) 576,8

7. (a) 0,881

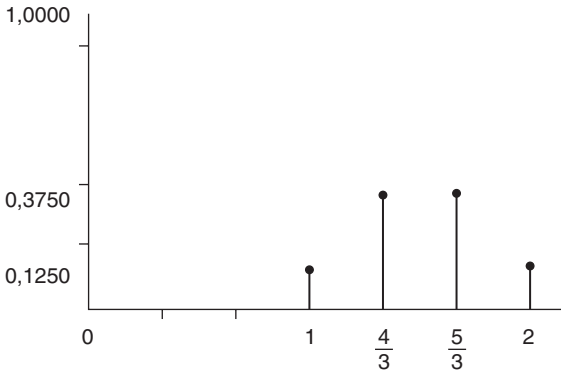
(b) 0,881

(c) 0,762

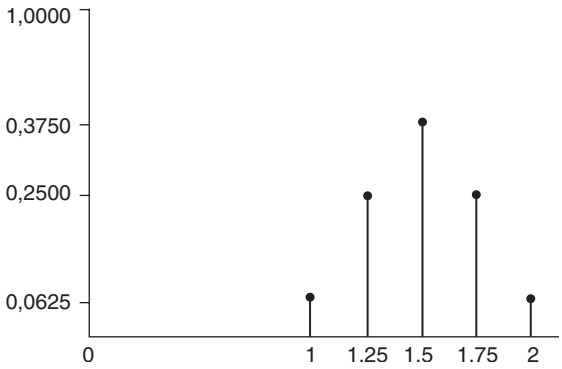
9. (a) 0,4483
   (b) 0,201
   (c) 0,4247
11. (a) 0,6915
   (b) 0,3859
   (c) 0,1587
13. (a) 1/4
   (b) 0,28965

Sección 7.3

1. (a) $\mathrm{SD}(\overline{X}) = \dfrac{1/2}{\sqrt{3}} \approx 0{,}29$



(b) $\mathrm{SD}(\overline{X}) = \dfrac{1/2}{\sqrt{4}} = 0{,}25$



3. (a) 2

   (b) $\sqrt{2/3} \approx 0{,}82$

   (c)

i	$P\{\overline{X} = i\}$
1	1/9
1,5	2/9
2	3/9
2,5	2/9
3	1/9

   (d) $E(\overline{X}) = 2$, $\mathrm{SD}(\overline{X}) = 1/\sqrt{3} \approx 0{,}58$

   (e) Sí

5. (a) $E(\overline{X}) = 2{,}4$, $\mathrm{SD}(\overline{X}) = 0{,}2/\sqrt{36} \approx 0{,}033$

   (b) $E(\overline{X}) = 2{,}4$, $\mathrm{SD}(\overline{X}) = 0{,}2/\sqrt{64} \approx 0{,}025$

   (c) $E(\overline{X}) = 2{,}4$, $\mathrm{SD}(\overline{X}) = 0{,}2/\sqrt{100} \approx 0{,}02$

   (d) $E(\overline{X}) = 2{,}4$, $\mathrm{SD}(\overline{X}) = 0{,}2/\sqrt{900} \approx 0{,}007$

7. El valor esperado es 15 500, y la desviación típica, 2800.

Sección 7.4

1. (a) 0,5468

   (b) 0,7888

   (c) 0,9876

3. 0,7888

5. (a) 0,0062

   (b) 0,7888

7. 0,9713

9. 0,1416

11. (a) 0,905

    (b) 0,5704

13. (a) 0

    (b) 0

15. (a) 0,6826
    (b) 0,9544
    (c) 1
    (d) 1
    (e) 1

Sección 7.5

1. (a) $E(\overline{X}) = 0{,}6, \text{SD}(\overline{X}) = 0{,}15$
   (b) $E(\overline{X}) = 0{,}6, \text{SD}(\overline{X}) = 0{,}049$
   (c) $E(\overline{X}) = 0{,}6, \text{SD}(\overline{X}) = 0{,}015$
   (d) $E(\overline{X}) = 0{,}6, \text{SD}(\overline{X}) = 0{,}0049$
3. (a) 0,0122
   (b) 0,119
   (c) 0,5222
9. (a) 0,0125
   (b) 0,8508
11. 0,1949
13. 0,4602
15. (a) 0,9147
    (b) 0,0043
    (c) 0,5188
17. (a) 0,9599
    (b) 0,3121
19. (a) 0,9974
    (b) 0,0268

Sección 7.6

1. (a) 5,7; 4 grados de libertad
   (b) 0,018; 5 grados de libertad
   (c) 1,13; 2 grados de libertad

Capítulo 7 Problemas de repaso

1. (a) 0,8413
   (b) 0,5
   (c) 0,0228
   (d) 0,0005
3. $E(\overline{X}) = 3; \text{SD}(\overline{X}) = 1/\sqrt{2} \approx 0{,}71$
5. (a) Media = 12, desviación típica = 3,25
   (b) 0,5588
7. (a) 300
   (b) $7\sqrt{2} \approx 31{,}3$
   (c) 0,5
9. 0,1003
11. (a) 0,3669
    (b) 0,9918
    (c) 0,9128

Sección 8.2

1. 145,5
5. 165,6 horas
7. 12
9. 3,23
11. (a)

Sección 8.3

1. 0,3849
3. 0,65; 0,107
5. 0,412; 0,05
7. (a) 0,122
   (b) 0,01

9. (a) 0,0233
   (b) 0,0375
   (c) 0,0867
11. (a) 0,245
   (b) 0,022
13. (c); precisión en los términos de menor error estándar.

Sección 8.3.1

1. 0,28
3. (b) 3,32; 1,73; 1,45

Sección 8.4

1. 18,36
3. 799,7; 193,12
5. 21,27
7. 30,5
9. 12,64
11. 1,35
13. 0,0474; 0,2386

Sección 8.5

1. (a) (3,06, 3,24)
   (b) (3,03, 3,27)
3. (11,43, 11,53)
5. (a) (8852,87, 9147,13)
   (b) (8824,69, 9175,31)
7. (72,53, 76,67)
9. (a) (1337,35, 1362,65)
   (b) (1334,92, 1365,08)
   (c) (1330,18, 1369,82)

11. 13,716
13. 3176
15. (a) 72,99
    (b) 72,53
    (c) 76,67
    (d) 77,53
17. No

Sección 8.6

1. (a) (5,15, 5,25)
   (b) (5,13, 5,27)
3. (a) (73,82, 93,91)
   (b) (71,63, 96,10)
   (c) (66,89, 100,84)
5. (a) (127,71, 163,29)
   (b) (119,18, 171,82)
7. (446,28, 482,01)
9. (280,04, 284,96)
11. (1849,4, 2550,6)
13. (a) (4,60, 4,80)
    (b) (4,58, 4,82)
15. (1124,95, 1315,05)
17. No
19. (a) (27,59, 38,64)
    (b) No
21. 68,897, 98,836
23. La media de las ventas de diarias superó los 2857 $.

Sección 8.7

1. (0,548, 0,660)
3. (a) (0,502, 0,519)
   (b) (0,498, 0,523)

5. (0,546, 0,734)
7. (0,359, 0,411)
9. (0, 0,306)
11. (0,801, 0,874)
13. (0, 0,45)
15. (a) (0,060, 0,108)
    (b) (0,020, 0,052)
    (c) (0,448, 0,536)
17. (a) Un intervalo al 95% de confianza viene dado por $0,75 \pm 0,0346$.
    (b) En lugar de utilizar $\hat{p}$ como estimador de p en el error estándar, se utilizó su cota superior $p(1 - p) \leq 1/4$.
19. (a) 1692
    (b) Menor que 0,04 pero mayor que 0,02.
    (c) (0,213, 0,247)
21. 6147
23. 0,868
25. (a) 0,139
    (b) 0,101
27. (a) No
    (b) No

Capítulo 8 Problemas de repaso

1. En el caso (a).
3. (22,35, 26,45)
5. (316,82, 323,18)
7. (a) (44,84, 54,36)
    (b) (45,66, 53,54)
9. (1527,47, 2152,53)
11. (a) 88,56
    (b) (83,05, 94,06)
13. (a) (34,02, 35,98)
    (b) (33,04, 36,96)
    (c) (31,08, 38,92)

15. (0,487, 0,549)

17. 0,004

19. (a) (0,373, 0,419)

    (b) (0,353, 0,427)

21. Superior

Sección 9.2

1. (a) Hipótesis *B*

3. (d) es la más correcta; puede decirse que (b) es más correcta que incorrecta.

Sección 9.3

1. TS = 1,55; $z_{\alpha/2} = 1{,}96$; no se rechaza H_0.

3. (a) 0,0026

   (b) 0,1336

   (c) 0,3174

   Al nivel de significación del 5%, se rechaza H_0 en (a). Al nivel de significación del 1%, se rechazar H_0 en (a).

5. Sí

7. (a) No

   (b) 0

9. Los datos no están a favor de que la media sea 13 500 millas.

11. Sí; sí.

13. El *p* valor es 0,281. En consecuencia, se rechaza esta hipótesis, al nivel de significación de 0,281 o a cualquier otro nivel mayor que éste.

15. (a) 0,2616

    (b) 0,2616

    (c) 0,7549

Sección 9.3.1

1. (a) No

   (b) No

   (c) 0,091

3. (a) 0

   (b) 0

   (c) 0,0085

5. (a) Sí

   (b) No, porque la reducción del número de caries es demasiado pequeña.

7. Sí, pero aumentando el tamaño muestral.

9. La cantidad media servida es inferior a 6 onzas; H_0: $\mu \geq 6$; H_1: $\mu < 6$; p valor $= 0$.

Sección 9.4

1. La evidencia no es lo suficientemente fuerte para refutar la tesis del productor, al nivel de significación del 5%.

3. (a) Sí

   (b) No

5. (a) No

   (b) No

   (c) No

   (d) El p valor es 0,108.

7. Sí

11. H_0: $\mu \geq 23$ frente a H_1: $\mu < 23$. La jueza debería dictaminar a favor de la panadería.

13. (a) H_0: $\mu \geq 31$

    (b) H_1: $\mu < 31$

    (c) No

    (d) No

15. No, el p valor es 0,0068.

17. No; no

Sección 9.5

1. p valor $= 0{,}0365$; con la aproximación normal se obtendría 0,0416.

3. No

5. (a) H_0: $p \leq 0{,}5$; H_1: $p > 0{,}5$

   (b) 0,1356

(c) 0,0519

(d) 0,0042

A medida que n crece, el p valor disminuye, puesto que la confianza en el estimador es mayor.

7. (a) No

(b) No

(c) No

(d) Sí

9. No; no

11. No

13. (a) Sí

(b) No

(c) 0,2005

Capítulo 9 Problemas de repaso

1. La (b)

5. (a) No

(b) Sí

(c) Sí

7. Existe una evidencia insuficiente a favor de la tesis mantenida, al nivel de significación del 5%.

9. Se debería dictaminar en contra de Caputo, puesto que el p valor, al contrastar H_0: $p = 1/2$ frente a H_1: $p \neq 1/2$ resultó ser 0,000016.

Sección 10.2

1. (a) No

(b) 0

3. (a) Existe evidencia a favor de que las longitudes medias de los cortes de las dos máquinas son iguales.

(b) 0,8336

5. Basta con cambiar un fichero de datos por otro y utilizar el mismo contraste.

7. No

Sección 10.3

1. Sí; H_0: $\mu_x = \mu_y$; H_1: $\mu_x \neq \mu_y$; p valor = 0,0206

3. p valor = 0,5664

5. Sí; 0

7. No; H_0: $\mu_x \leq \mu_y$; H_1: $\mu_x > \mu_y$, donde x se corresponde con los estudiantes rurales, e y con los estudiantes urbanos.

9. H_0: $\mu_B \leq \mu_A$; H_1: $\mu_B > \mu_A$; p valor = 0,0838. Al nivel de significación del 5%, se debería elegir al suministrador B.

11. (a) H_0: $\mu_m \leq \mu_f$; H_1: $\mu_f < \mu_m$

    (c) Indica que el salario medio de las mujeres es inferior al salario medio de los hombres.

13. (a) Se debería rechazar la hipótesis nula para $\alpha = 0{,}01$.

    (b) 0,0066

    (c) Produjo una reducción en la puntuación media.

Sección 10.4

1. No; sí

3. (a) No

    (b) No

5. Sí

7. Se debe rechazar H_0: $\mu_x = \mu_y$ para $\alpha = 0{,}05$. p valor = 0,0028.

9. (a) Se rechaza H_0: $\mu_x = \mu_y$

    (b) Se rechaza H_0: $\mu_x = \mu_y$

    (c) No se rechaza H_0: $\mu_x = \mu_y$

Sección 10.5

1. (a) Se rechaza la hipótesis, al nivel $\alpha = 0{,}05$.

    (b) p valor = 0,0015

3. No se rechaza H_0.

5. (a) No se rechaza la hipótesis.

    (b) No existe evidencia para rechazar la hipótesis, al nivel de significación del 5%.

7. Se rechaza la hipótesis, al nivel $\alpha = 0{,}05$.
9. (a) H_0: $\mu_{\text{antes}} \leq \mu_{\text{después}}$; H_1: $\mu_{\text{antes}} > \mu_{\text{después}}$
   (b) No
11. No se rechaza la hipótesis nula.

Sección 10.6

1. (a) No
   (b) No
3. (a) Sí
   (b) 0,0178
5. (a) No
   (b) 0,0856
7. Se rechaza la hipótesis de que las proporciones de 1983 y 1990 coincidían; p valor = 0,0017.
9. Se rechaza la hipótesis, al nivel $\alpha = 0{,}05$; p valor = 0.
11. (a) Sí
    (b) 0
13. No
15. Sí; H_0: $\hat{p}_{\text{placebo}} \leq \hat{p}_{\text{aspirin}}$(donde $\hat{p}$ representa la proporción muestral de físicos que sufrieron ataques de corazón); $\hat{p}$ valor = 0.

Capítulo 10 Problemas de repaso

1. (a) Se rechaza H_0: $\mu_x = \mu_y$
   (b) 0
3. (a) No se rechaza la hipótesis de que las probabilidades son iguales.
   (b) 0,5222
   (c) No
   (d) $\alpha \geq 0{,}2611$
5. (a) Se rechaza H_0: $\mu_x = \mu_y$
   (b) 0,0497
7. No se rechaza la hipótesis de que las probabilidades son iguales.

9. Se rechaza la hipótesis (p valor $= 0{,}79$).

11. No se rechaza la hipótesis de que las proporciones sean iguales en ambos deportes.

Sección 11.2

1. (a) $\overline{X}_1 = 8, \overline{X}_2 = 14, \overline{X}_3 = 11$

   (b) $\overline{\overline{X}} = 11$

3. Sí

5. No

7. No se rechaza la hipótesis, al nivel $\alpha = 0{,}05$.

9. Se rechaza la hipótesis de que las tasas de mortalidad no dependen de la estación, para $\alpha = 0{,}05$.

11. No

Sección 11.3

1. $\hat{\alpha} = 68{,}8, \hat{\alpha}_1 = 14{,}2, \hat{\alpha}_2 = 6{,}53, \hat{\alpha}_3 = -3{,}47, \hat{\alpha}_4 = -3{,}47, \hat{\alpha}_5 = -13{,}8, \hat{\beta}_1 = 0{,}8, \hat{\beta}_2 = -2{,}4, \beta_3 = 1{,}6$

3. $\hat{\alpha} = 28{,}33, \hat{\alpha}_1 = 1, \hat{\alpha}_2 = -2, \hat{\alpha}_3 = 1, \hat{\beta}_1 = 3{,}67, \hat{\beta}_2 = -0{,}67, \hat{\beta}_3 = -3$

7. $\hat{\alpha} = 9{,}58, \hat{\alpha}_1 = -1{,}74, \hat{\alpha}_2 = -1{,}96, \hat{\alpha}_3 = 4{,}915, \hat{\alpha}_3 = 4{,}915, \hat{\alpha}_4 = -1{,}36, \hat{\alpha}_5 = -3{,}335, \hat{\beta}_1 = 0{,}495, \hat{\beta}_2 = -0{,}405, \hat{\beta}_3 = 0{,}795, \hat{\beta}_4 = -0{,}885$

9. (a) 44

   (b) 48

   (c) 52

   (d) 144

Sección 11.4

1. (a) Sí

   (b) No

3. (a) No

   (b) No

5. (a) No (H_0 se rechaza)

   (b) Sí (H_0 no se rechaza)

9. (a) Se rechaza la hipótesis, al nivel $\alpha = 0{,}05$.

   (b) No se rechaza la hipótesis, para $\alpha = 0{,}05$.

Capítulo 11 Problemas de repaso

1. Se rechaza la hipótesis, al nivel $\alpha = 0{,}05$.

3. Sí, para $\alpha = 0{,}05$.

5. No se rechaza la hipótesis, para $\alpha = 0{,}05$.

7. (a) No se rechaza la hipótesis, al nivel $\alpha = 0{,}05$.

   (b) 30,6

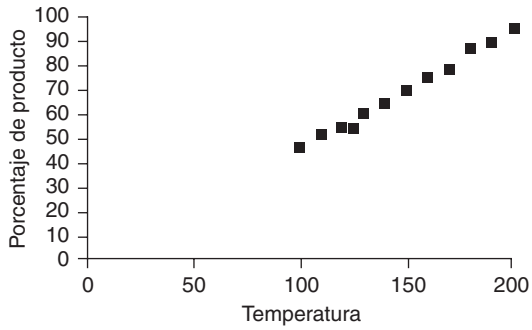
   (c) Se rechaza la hipótesis, para $\alpha = 0{,}05$.

9. (a) No se rechaza la hipótesis, al nivel $\alpha = 0{,}05$.

   (b) Se rechaza la hipótesis, para $\alpha = 0{,}05$.

Sección 12.2

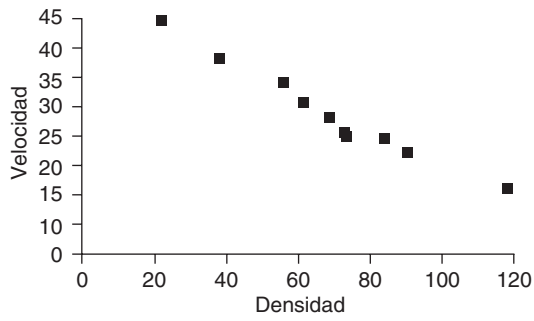
1. (a)



   (b) Sí

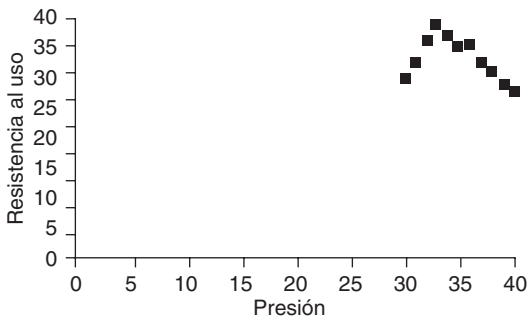
3. (a) Densidad; velocidad

   (b)



   (c) Sí

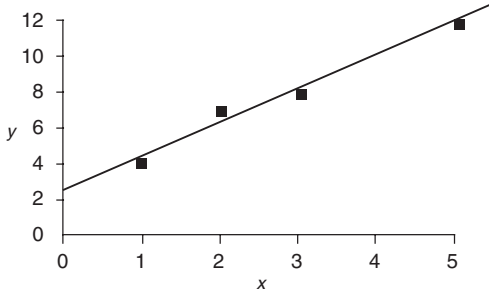
5. (a)



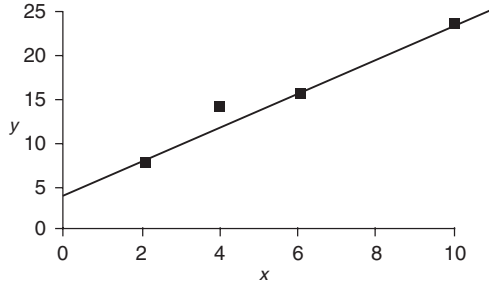
(b) No

Sección 12.3

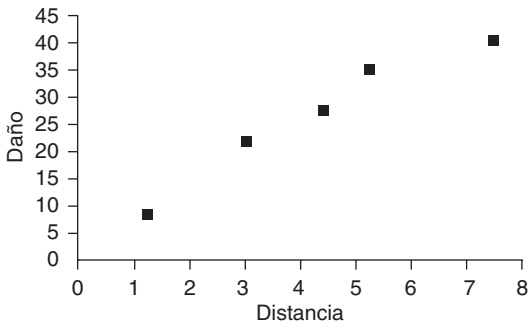
1. (a)



(b)



3. (a)



(c) $y = 14{,}79 + 2{,}43x$

7. (a) $y = -8{,}31 + 0{,}27x$

   (b) 31,66

   (c) $y = 31{,}66 + 3{,}61x$

   (d) 147,12

9. Aleatoriamente

11. (a) $y = 67{,}56 + 0{,}23x$

    (b) 204,62

    (c) 261,73

    (d) 296,00

13. 121,85

Sección 12.4

1. 2,32

3. (a) 6

   (b) 6

   (c) 76

5. 0,000156

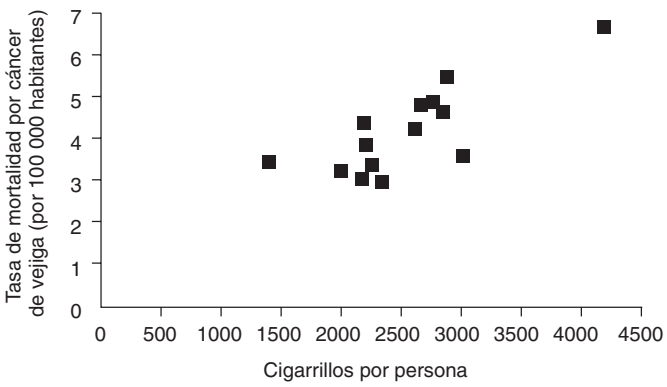
7. 6970,21

Sección 12.5

1. No se rechaza H_0: $\beta = 0$.

3. Se rechaza la hipótesis.

5. (a)

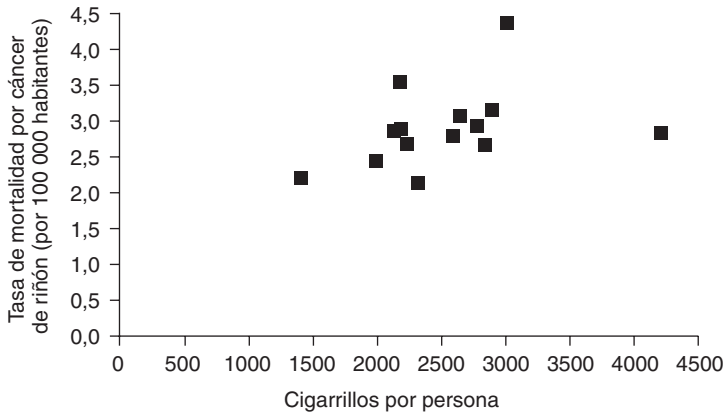


(b) $y = 0{,}75 + 0{,}0013x$

(c) Se rechaza la hipótesis.

(d) Se rechaza la hipótesis.

7. (a)



(b) $y = 2{,}12 + 0{,}0003x$

(c) No se rechaza la hipótesis.

(d) No se rechaza la hipótesis.

9. Se rechaza la hipótesis.

11. (a) No se rechaza la hipótesis, al nivel $\alpha = 0{,}05$.

(b) No se rechaza la hipótesis, al nivel $\alpha = 0{,}05$.

(c) No se rechaza la hipótesis, al nivel $\alpha = 0{,}05$.

Sección 12.6

1. (a) $\alpha = 10{,}48$, $\beta = 0{,}325$

(b) Sí

7. El ajuste no es tan bueno como con las alturas.

Sección 12.7

1. (a) 12,6

(b) (6,4, 18,8)

3. (a) $y = 45{,}42 - 0{,}66x$

(b) 7773,49 $

(c) (6659,78, 8887,21)

(d) (6026,89, 9520,09)

5. (a) 2,501

   (b) (2,493, 2,510)

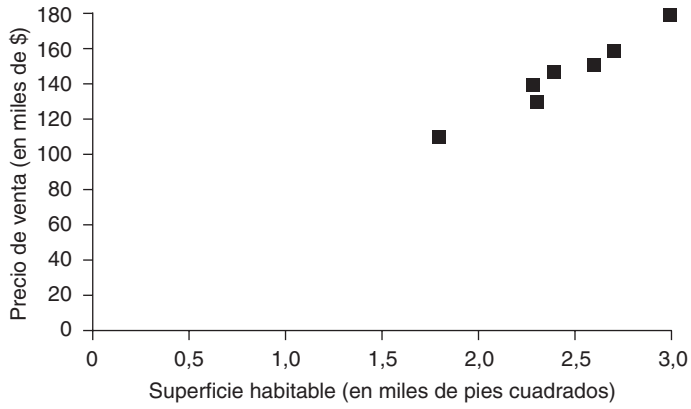
7. (a) 33 266 $

   (b) (27 263, 39 268)

   (c) 42 074 $; (35 608, 48 541)

Sección 12.8

1. (a)



   (b) $y = 8{,}885 + 56{,}32x$

   (c) 97%

   (d) (144,628, 165,929)

3. (a) 0,9996

   (b) Sí

   (c) 41,975

   (d) (40,692, 43,258)

5. 0,149

7. 0,059

Sección 12.9

1. (a) 0,9796; 0,9897

   (b) 0,9796; 0,9897

   Esto indica que el valor del coeficiente de correlación muestral no depende de qué variable independiente se tome.

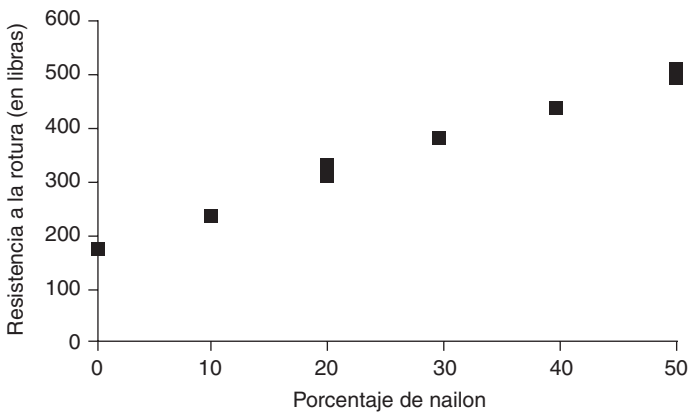
3. (a) 0,8
   (b) 0,8
   (c) −0,8
   (d) −0,8
5. (a) $y = -3{,}16 + 1{,}24x$
   (b) $y = 7{,}25 + 0{,}66x$
   (c) 0,818; 0,904
   (d) 0,818; 0,904

Sección 12.11

3. $y = -153{,}51 + 51{,}75x_1 + 0{,}077x_2 + 20{,}92x_3 + 13{,}10x_4$; 183,62
5. 69,99

Capítulo 12 Problemas de repaso

1. (a)



   (b) $y = 177{,}93 + 6{,}89x$
   (c) 522,61
   (d) (480,53, 564,68)
3. (a) $\alpha = 94{,}13$; $\beta = 0{,}155$
   (b) (93,17, 132,34)
   (c) 100%
5. No necesariamente, los aterrizajes buenos (o malos) podrían ser simplemente un fenómeno debido al azar que tenderán a regresionar a la media en el siguiente intento.

9. (a) 34,9

   (b) (4,34, 23,40)

11. (a) $y = 177{,}41 + 1{,}07x_1 + 11{,}7x_2$

    (b) 241,90

15. El consumo de alcohol, que está asociado tanto con el consumo de cigarrillos como con el cáncer de vejiga, podría ser la causa primaria. Una regresión lineal múltiple podría ser útil.

Sección 13.2

1. (a) 15,086

   (b) 11,070

   (c) 23,209

   (d) 18,307

   (e) 31,410

3. H_0: $P_1 = 0{,}52$, $P_2 = 0{,}32$, $P_3 = 0{,}16$. No, no se rechaza H_0.

5. Sí, la hipótesis nula resulta rechazada.

7. No, p valor $= 0{,}0002$

9. Sí; sí

11. Sí

13. No se rechaza la hipótesis.

15. Se rechaza la hipótesis.

Sección 13.3

1. (a) 7,08

   (b) Sí

   (c) No

3. Se rechaza la hipótesis.

5. No se rechaza que las características son independientes.

7. Se rechaza la hipótesis.

9. No

11. Se rechaza la hipótesis; se rechaza la hipótesis.

13. No se rechaza la hipótesis.

Sección 13.4

1. No, no se puede concluir que fumar sea la causa del cáncer de pulmón; sin embargo, sí se puede decir que la tasa de cáncer de pulmón es mayor para los fumadores que para los no fumadores.

3. No se rechaza la hipótesis.

5. No

7. Sí; no

9. No se rechaza en ninguno de los casos.

Capítulo 13 Problemas de repaso

1. No se rechaza la hipótesis.

5. Se rechaza la hipótesis.

7. No se rechaza la hipótesis, al nivel $\alpha = 0{,}05$.

9. Sí

11. No; no

13. (a) No se rechaza la hipótesis.

    (b) 0,208

15. No se rechaza la hipótesis; no se rechaza la hipótesis.

Sección 14.2

1. (a) p valor $= 0{,}057$. Se rechaza la hipótesis nula a cualquier nivel de significación mayor o igual que 0,057.

   (b) p valor ≈ 0. Se rechaza la hipótesis nula a cualquier nivel de significación.

   (c) p valor ≈ 0. Se rechaza la hipótesis nula a cualquier nivel de significación.

3. No se puede rechazar la hipótesis nula de que las dos pistolas sean igualmente efectivas.

5. Puesto que n es pequeño, se debe utilizar la distribución binomial para calcular el p valor $= 0{,}291$. Así pues, no se puede rechazar la hipótesis de que la mediana de las calificaciones es al menos de 72 puntos.

7. Sí, está en contra de la hipótesis. p valor $= 0{,}0028$.

Sección 14.3

1. (a) TS = 39

   (b) TS = 42

   (c) TS = 20

3. (a) p valor = 0,2460

   (b) p valor = 0,8336

   (c) p valor = 0,1470

5. (a) Sí, la forma como se presenta el trabajo influye en la calificación obtenida.

   (b) p valor = 0,0102

7. (a) Se rechaza la hipótesis nula a cualquier nivel de significación mayor o igual que 0,1250.

   (b) Se rechaza la hipótesis nula a cualquier nivel de significación mayor o igual que 0,0348.

9. No, no se puede rechazar la hipótesis nula. La pintura no afecta a la velocidad de crucero de las avionetas.

Sección 14.4

1. (a) 94

   (b) 77

3. p valor = 0,8572

5. Puesto que el p valor = 0,2112, no se puede rechazar la hipótesis nula de que las distribuciones de los salarios iniciales de los graduados por las dos universidades sean iguales.

7. p valor = 0,4357

Sección 14.5

1. (a) 41

   (b) 2

3. Puesto que el p valor = 0,0648, no se puede rechazar la hipótesis nula de que los datos constituyan una muestra aleatoria.

5. Puesto que el p valor = 0,0548, no se puede rechazar la hipótesis nula de que el entrevistador les haya encuestado siguiendo un orden elegido aleatoriamente.

7. (a) Mediana = 163,5

   (b) Siete rachas.

   (c) Puesto que el p valor = 0,0016, se debe rechazar la hipótesis nula a cualquier nivel de significación mayor o igual que 0,0016. La sucesión de valores no constituye una muestra aleatoria.

Capítulo 14 Problemas de repaso

1. Si se utiliza el contraste de la suma de rangos con TS = 113, se obtiene un p valor de 0,0348. Así pues, no se puede rechazar la hipótesis nula, al nivel de significación del 1%; sin embargo, sí se rechazaría esta hipótesis al nivel del 5%.

3. Puesto que el p valor ≈ 0, se rechaza la hipótesis nula; el salario neto mediano ha disminuido.

5. Se lleva a cabo un contraste de rachas, con la mediana = 145, $n = m = 20$ y $r = 21$. Puesto que $\mu = 21$, el p valor es 1,0.

9. Se usa el contraste de rangos signados con TS = 0. El p valor = 0,044. Así pues, se rechaza la hipótesis nula de que no existan diferencias en las ventas de zapatos, a cualquier nivel de significación por encima del 4,44%.

11. Puesto que el p valor = 0,5620, no se puede rechazar la hipótesis nula.

Sección 15.2

1. (a) LCL = 85, UCL = 115

   (b) LCL = 86,58, UCL = 113,42

   (c) LCL = 87,75, UCL = 112,25

   (d) LCL = 90,51, UCL = 109,49

3. LCL = 66,58, UCL = 93,42. Puesto que el subgrupo 9 cae fuera de este rango, el proceso se debería haber declarado fuera de control en ese punto.

5. LCL = −0,00671, UCL = 0,00671. Puesto que todos los subgrupos están dentro de estos límites de control, el proceso ha estado bajo control.

Sección 15.3

1. LCL = 0, UCL = 13,23. Puesto que todos los subgrupos están dentro de los límites de control, el proceso ha estado bajo control.

3. (a) Puesto que todos los subgrupos están dentro de los límites de control, el proceso se ha mantenido bajo control.

   (b) LCL = 0, UCL = 9,88

Capítulo 15 Problemas de repaso

1. LCL = 1,4985, UCL = 1,5015.

3. LCL = 0, UCL = 13,23. Puesto que todos los subgrupos están dentro de estos límites de control, el proceso ha estado bajo control.

Índice alfabético

La "f" y la "t" adjuntas al número de página denotan una figura o una tabla, respectivamente.

Contrastes de hipótesis relativos a las proporciones de dos poblaciones

p_1 y p_2 son las proporciones de los miembros de las dos poblaciones que presentan una determinada característica. Se extraen una muestra de tamaño n_1 de la primera población y otra muestra independiente de tamaño n_2 de la segunda población. $\hat{p}_1$ y $\hat{p}_2$ son las proporciones muestrales de individuos que presentan la característica y $\hat{p}$ es la proporción de individuos que la presentan en la muestra combinada.

H_0	H_1	Estadístico del contraste TS	Contraste a nivel de significación α	p valor si TS $= v$
$p_1 = p_2$	$p_1 \neq p_2$	$\dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{(1/n_1 + 1/n_2)\hat{p}(1-\hat{p})}}$	Rechazar H_0 si $\lvert TS\rvert \geq z_{\alpha/2}$	$2P\{Z \geq \lvert v\rvert\}$
$p_1 \leq p_2$	$p_1 > p_2$	$\dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{(1/n_1 + 1/n_2)\hat{p}(1-\hat{p})}}$	Rechazar H_0 si TS $\geq z$	$P\{Z \geq v\}$

11 Análisis de la varianza

Tabla ANOVA unifactorial

$\bar{X}_i$ y S_i^2, $i = 1, \ldots, m$, son las medias muestrales y las varianzas muestrales de muestras independientes de tamaño n procedentes de poblaciones normales con media μ_i y varianzas iguales σ^2.

Fuente del estimador	Estimador de σ^2	Valor del estadístico del contraste
Entre las muestras	$n\hat{S}^2 = \dfrac{n\sum_{i=1}^{m}(\bar{X}_i - \bar{\bar{X}})^2}{(m-1)}$	$\text{TS} = \dfrac{n\hat{S}^2}{\left(\sum_{i=1}^{m} S_i^2\right)\Big/m}$
Dentro de las muestras	$\left(\sum_{i=1}^{m} S_i^2\right)\Big/m$	

Contraste a nivel de significación α de H_0: todos los μ_i son iguales
Rechazar H_0 si TS $\geq F_{m-1,\, m(n-1),\alpha}$
No rechazar en caso contrario

Si TS $= v$ se tiene que

$$p \text{ valor } = P\{F_{m-1,\, m(n-1)} \geq v\}$$

donde $F_{m-1,\, m(n-1)}$ es una variable aleatoria F con $m-1$ grados de libertad en el numerador y $m(n-1)$ grados de libertad en el denominador.

Modelo ANOVA bifactorial: Para $i = 1, \ldots, m$, $j = 1, \ldots, n$

$$E[X_{ij}] = \mu + \alpha_i + \beta_j$$
$$\sum_{i=1}^{m}\alpha_i = \sum_{j=1}^{n}\beta_j = 0$$

μ es la media total, α_i es la desviación de la media total debida a la fila i y β_j es la desviación de la media total debida a la columna j. Sus estimadores son

$$\hat{\mu} = X_{..} \qquad \hat{\alpha}_i = X_{i.} - X_{..} \qquad \hat{\beta}_j = X_{.j} - X_{..}$$

Tabla ANOVA bifactorial

	Suma de cuadrados	Grados de libertad
Filas	$\text{SS}_r = n\sum_{i=1}^{m}(X_{i.} - X_{..})^2$	$m-1$
Columnas	$\text{SS}_c = m\sum_{j=1}^{n}(X_{.j} - X_{..})^2$	$n-1$
Error	$\text{SS}_e = \sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} - X_{i.} - X_{.j} + X_{..})^2$	$N = (n-1)(m-1)$

Hipótesis nula	Estadístico del contraste	Contraste a nivel de significación α	p valor si TS $= v$
No existe efecto fila (todos los $\alpha_i = 0$)	$\dfrac{\text{SS}_r/(m-1)}{\text{SS}_e/N}$	Rechazar H_0 si TS $\geq F_{n-1,N,\alpha}$	$P\{F_{m-1,N} \geq v\}$
No existe efecto columna (todos los $\beta_j = 0$)	$\dfrac{\text{SS}_c/(n-1)}{\text{SS}_e/N}$	Rechazar H_0 si TS $\geq F_{n-1,N,\alpha}$	$P\{F_{n-1,N} \geq v\}$

12 Regresión lineal

Modelo de regresión lineal simple: $Y = \alpha + \beta x + e$
Estimadores por mínimos cuadrados: $\hat{\beta} = S_{xy}/S_{xx}$, $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y}) = \sum_{1}^{n} x_i Y_i - n\bar{x}\bar{Y}$$

$$S_{xx} = \sum_{1}^{n}(x_i - \bar{x})^2 = \sum_{1}^{n} x_i^2 - n\bar{x}^2$$

Recta de regresión estimada: $y = \hat{\alpha} + \hat{\beta}x$

Copyright © 2005 by Elsevier, Inc. All rights reserved.

El término de error e es normal con media 0 y varianza σ^2. El estimador de σ^2 es $SS_R/(n-2)$, $SS_R = \sum_i (Y_i - \hat{\alpha} - \hat{\beta}x_i)^2 = (S_{xx}S_{YY} - S_{xY}^2)/S_{xx}$

Para contrastar H_0: $\beta = 0$. Utilice $TS = \sqrt{(n-2)S_{xx}/SS_R}\,\hat{\beta}$
El contraste a nivel de significación γ consiste en rechazar H_0 si $|TS| \geq t_{n-2,\,\gamma/2}$.
Si $TS = v$, p valor $= 2P\{T_{n-2} \geq v\}$
El intervalo de predicción, a confianza de $100(1-\gamma)$, para la respuesta al valor de entrada x_0 es

$$\hat{\alpha} + \hat{\beta}x_0 \pm t_{n-2,\gamma/2}\sqrt{(1 + 1/n + (x_0 - \bar{x})^2/S_{xx})SS_R/(n-2)}$$

Coeficiente de determinación: $R^2 = 1 - SS_R/S_{YY}$ es la proporción de la variación en las variables de respuesta que es explicada por los diferentes valores de entrada. Su raíz cuadrada coincide con el valor absoluto del coeficiente de correlación muestral.

Modelo de regresión múltiple:
$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + e$

13 Contrastes de bondad de ajuste de la chi-cuadrado

P_i es la proporción de elementos de la población con valor i, $i = 1, \ldots, k$.

Para contrastar H_0: $P_i = p_i$, $i = 1, \ldots, k$, extraiga una muestra de tamaño n. Sea N_i el número de elementos de la muestra con valor i, $e_i = np_i$, $TS = \sum_{i=1}^{k}(N_i - e_i)^2/e_i$. El contraste, a nivel α rechaza H_0 si $TS \geq \chi^2_{k-1,\alpha}$.

Si $TS = v$, se tiene que p valor $= P\{\chi^2_{k-1} \geq v\}$.

Supongamos que cada miembro de la población presenta una característica X y otra Y. Supongamos que existen r posibles valores de la característica X y s posibles valores de Y. Para contrastar la independencia de las características de un miembro elegido aleatoriamente extraiga una muestra de tamaño n.

N_{ij} = número de elementos de la muestra con el valor i para la característica X y con el valor j para la característica Y.
N_i = número de elementos de la muestra con el valor i para la característica X
M_j = número de elementos de la muestra con el valor j para la característica Y

$\hat{e}_{ij} = N_i M_j/n$

Si $\sum_i \sum_j (N_{ij} - \hat{e}_{ij})^2/\hat{e}_{ij} \geq \chi^2_{(r-1)(s-1),\alpha}$ la hipótesis de independencia debe ser rechazada al nivel de significación α.

14 Hipótesis no paramétricas

Sea η = mediana de la población. El contraste de *signos* para

H_0: $\eta = m$ frente a H_1: $\eta \neq m$

comienza extrayendo una muestra de tamaño n. Si i es igual al número elementos de la muestra con valores menores que m, se tiene que

p valor $= 2$ Min $(P\{N \leq i\}, P\{N \geq i\})$

donde N representa una variable aleatoria binomial $(n, 1/2)$.

El contraste de *rangos signados* se utiliza para contrastar la hipótesis de que la distribución poblacional es simétrica respecto de 0. Ordena los datos en términos de sus valores absolutos. TS es la suma de los rangos de los valores negativos. Si $TS = t$, se tiene que

p valor $= 2$ Min $(P\{TS \leq t\}, P\{TS \geq t\})$

TS sigue aproximadamente una normal con media $n(n+1)/4$ y varianza $n(n+1)(2n+1)/24$.

Para contrastar la igualdad de dos distribuciones poblacionales, extraiga de cada una de las poblaciones dos muestras de tamaños n y m y ordene los $n + m$ valores de datos. El contraste de la *suma de rangos* utiliza TS = suma de los rangos de la primera muestra. Se rechaza H_0 si TS es significativamente grande o bien es significativamente pequeño. Si $TS = t$, se tiene que

p valor $= 2$ Min $(P\{TS \leq t\}, P\{TS \geq t\})$

TS sigue aproximadamente una normal con media $n(n+m+1)/2$ y varianza $nm(n+m+1)/12$.

Para contrastar la hipótesis de que una sucesión de ceros y unos es aleatoria, utilice el contraste de rachas que precisa contabilizar R, el número de *rachas*. La aleatoriedad se rechazará si R es demasiado pequeño o bien si es demasiado grande como para poder ser atribuido al azar. Utilice el resultado de que, cuando H_0 es cierta, R sigue aproximadamente una normal con media $1 + 2nm/(n+m)$ y varianza

$$\frac{2nm(2nm - n - m)}{(n+m)^2(n+m-1)}$$

15 Control de calidad

Límites de los gráficos de control $\mu \pm 3\sigma/\sqrt{n}$ n = tamaño del subgrupo

Área que queda a la izquierda de x bajo la curva de la normal estándar

x	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5733
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998

Copyright © 2005 by Elsevier, Inc. All rights reserved.

Sheldon M. Ross

Introducción a la **Estadística**

Hoy en día vivimos en un mundo repleto de información, es decir, de datos, y no cabe duda que para interpretarlos correctamente es fundamental el conocimiento de la Estadística, que podría definirse como el arte de obtener conclusiones a partir de datos.

Dirigida a estudiantes universitarios de cualquier área, esta Introducción a la Estadística sólo exige conocimientos de álgebra a nivel de enseñanza media. Su objetivo no es simplemente el de presentar conceptos y técnicas estadísticas, sino que pretende que los futuros profesionales sepan cuándo y cómo deben aplicar los procedimientos estadísticos y, además, entiendan la razón por la cual se utiliza uno en concreto en determinados casos. Para ello, los autores han realizado un gran esfuerzo a la hora de explicar las ideas que sustentan los conceptos y las técnicas estadísticas presentadas.

Las aplicaciones de la Estadística y las distintas perspectivas de su uso se explican aquí de forma clara y concisa, y se ilustran con numerosos ejemplos y problemas de trabajo sobre una amplia variedad de temas, en su mayor parte tomados de la vida real. Para desarrollar las habilidades del lector, se proponen cientos de ejercicios y problemas de repaso que incitan a pensar. En www.reverte.com se puede descargar el programa STATCOMP, una herramienta muy útil a la hora de resolver estos ejercicios.

EDITORIAL REVERTÉ
www.reverte.com

Publicado en su versión original con el título **Introductory Statistics,** Second Edition. Traducido del inglés con la autorización de ediciones Elsevier.

ELSEVIER

Cover Image: © Getty Images